Introduction à l'optimisation

Mondher FARZA

Université de Caen, ENSICAEN
mondher.farza@unicaen.fr

CHAPITRE 1

PROBLEMES D'OPTIMISATION DANS $\mathbb{R}^n$

Formulation du problème d'optimisation

Un problème d'optimisation consiste à trouver un vecteur $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ qui

minimise $f(x)$ sous les contraintes

- de type inégalités: $g_j(x) \leq 0$ pour $j = 1, \ldots, m$.
- de type égalités: $l_j(x) = 0$ pour $j = 1, \ldots, p$.

En général, il n'y a pas de lien entre le nombre de variables n et le nombre des contraintes $m + p$. Lorsque $m + p = 0$, le problème est dit problème d'optimisation sans contrainte.

• **variables de synthèse - variables décisionnelles:**

Tout système peut être décrit par un ensemble de quantités parmi lesquelles certaines peuvent varier au cours du processus de synthèse et d'autres sont fixées ou imposées par les conditions environnementales. Toutes les quantités qui peuvent varier sont appelées variables de synthèse ou encore variables décisionnelles et sont rassemblées au sein du vecteur x.

• **Contraintes de conception :**

En pratique, les variables de conception ne peuvent pas être choisies arbitrairement, mais doivent satisfaire certaines exigences. Ces restrictions sont appelées des contraintes de conception. Les contraintes de conception peuvent présenter des limitations sur la performance ou le comportement du système.

• **Fonction objectif:**

Dans une procédure classique de conception, on cherche la meilleure conception possible qui satisfasse aux contraintes de conception. En général, plusieurs conceptions satisfaisant aux contraintes sont possibles et le but de l'optimisation consiste alors à chercher parmi ces conceptions la meilleure. De ce fait, un critère devrait être choisi pour comparer les différentes conceptions. Lorsque ce critère est exprimé comme fonction des variables de conception, on l'appelle fonction objectif. La spécification de la fonction objectif tient en général compte de considérations physiques ou économiques. Cependant, une telle spécification n'est pas en général triviale parce qu'une solution optimale pour un certain critère peut être inacceptable pour un autre critère. En général, il y a un compromis entre performance et coût, ou performance et faisabilité.

Classification des problèmes d'optimisation

Les problèmes d'optimisation peuvent être classés de plusieurs façons.

• Existence de contraintes: problèmes avec ou sans contraines.

• Nature des équations: les problèmes d'optimisation peuvent être qualifiés de linéaire, quadratique, polynômial ou non linéaire selon la nature de la fonction objectif et des contraintes. Une telle classification est importante car d'elle dépend la selection des méthodes de résolution du problème d'optimisation considéré.

• Valeurs admissibles des variables décisionnelles: selon les valeurs que peuvent prendre les variables décisionnelles, les problèmes d'optimisation peuvent être classés en tant que problème d'optimisation à valeurs réelles ou à valeurs entières ou encore problème d'optimisation déterministe ou stochastique.

Exemple: Minimisation au sens des moindres carrées

On dispose d'un modèle paramétrique $z = f(x, y)$ décrivant les variations d'une grandeur réelle z en fonction d'une ou de plusieurs grandeurs rassemblées au sein du vecteur y et d'un vecteur de paramètres inconnus x. En supposant connu un échantillon (z_1, y_1), (z_2, y_2), ..., (z_M, y_M) de couples de valeurs de z et de y, l'objectif est d'estimer le vecteur des paramètres $x \in \mathbb{R}^N$ en minimisant la somme des carrés entre les valeurs observées et les valeurs prévues par le modèle:

$$S(x) = \sum_{j=1}^{M} |z_j - f(x, y_j)|^2 \quad (\star)$$

• **Exemple 1:**

La loi logistique :$z = \dfrac{a}{b + ce^{-at}}$ fournit un modèle pour calculer la population mondiale en fonction du temps t. Le vecteur des paramètres est $x = (a, b, c)^T$. En supposant que le nombre de la population mondiale, z_j, est connu à l'instant t_j pour $j = 1, \ldots, M$, le critère $(\star)$ devient dans ce cas:

$$S(a, b, c) = \sum_{j=1}^{M} \left| z_j - \frac{a}{b + ce^{-at_j}} \right|^2$$

• **Exemple 2:**

La loi de Monod : $\mu = \dfrac{\mu_{max} C}{K_C + C}$ fournit un modèle pour le taux spécifique de croissance d'une biomasse consommant un substrat C. Le vecteur des paramètres est $x = (\mu_{max}, K_C)^T$ où μ_{max} est le taux spécifique maximum de croissance et K_C est la constante de saturation. En supposant qu'un échantillon de mesures $(\mu_1, C_1), (\mu_2, C_2), \ldots, (\mu_M, C_M)$ soit disponible, le critère $(\star)$ s'écrit dans ce cas comme suit:

$$J(\mu_{max}, K_C) = \sum_{j=1}^{M} \left| \mu_j - \frac{\mu_{max} C_j}{K_C + C_j} \right|^2$$

• **Exemple 3:**

On reprend la la même loi de Monod mais maintenant, on ne dipose plus de mesures de μ mais plutôt du modèle dynamique suivant:

$$\begin{cases} \dot{C} = -k_1 \dfrac{\mu_{max} C}{K_C + C} X \\ \dot{X} = \dfrac{\mu_{max} C}{K_C + C} X \end{cases}$$

où C et X sont respectivement le substrat et la biomasse qui sont mesurés à chaque instant.

CHAPITRE 2

DERIVABILITE/DIFFERENTIABILITES DE FONCTIONS

Dérivée de fonctions de $\mathbb{R}$ dans $\mathbb{R}^m$:

Soit I un intervalle ouvert de $\mathbb{R}$ et $f = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix} : I \to \mathbb{R}^m$, où

$f_i : I \to \mathbb{R}$ pour $i = 1, \ldots, m$. Soit $x^\star$ un élément de I.

Définition 1 *On dit que la fonction f_i, $i \in \{1, \ldots, m\}$, est dérivable en $x^\star$ si la quantité suivante existe:*

$$\lim_{h \to 0} \frac{f_i(x^\star + h) - f_i(x^\star)}{h}$$

Dans ce cas, cette limite est appelée la dérivée de f_i en $x^\star$ et elle est notée $f_i'(x^\star)$.

Lemme 1 *La fonction f est dérivable en $x^\star$ si et seulement si chaque fonction f_i, $i = 1, \ldots, m$ est dérivable en $x^\star$. De plus, on a:*

$$f'(x^\star) = \begin{pmatrix} f_1'(x^\star) \\ f_2'(x^\star) \\ \vdots \\ f_m'(x^\star) \end{pmatrix}$$

Dérivée de fonctions de $\mathbb{R}^n$ dans $\mathbb{R}^m$:

Soit Ω un ouvert de $\mathbb{R}^n$ et $f\colon \Omega \to \mathbb{R}^m$, $f = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix}$ où $f_i\colon \Omega \to \mathbb{R}$,

$i = 1, \ldots, m$.

La définition précédente n'a plus de sens car on ne peut pas diviser par un vecteur. Ceci nous amène à construire à partir de f des fonctions d'une variable scalaire.

• **Dérivée suivant un vecteur:**

Soit Ω un ouvert de $\mathbb{R}^n$, $x^\star$ et u deux éléments de Ω, $u \neq 0$, et soit l'ensemble $I = \{t \in \mathbb{R} / x^\star + tu \in \Omega\}$. On définit la fonction d'une variable scalaire suivante $\varphi : I \to \mathbb{R}^m$, $t \mapsto \varphi(t) = f(x^\star + tu)$.

Définition 2 *L'application f est dérivable au point $x^\star$ dans la direction de u si la fonction φ est dérivable en 0. Dans ce cas, $\varphi'(0)$ est appelée la dérivée de f selon le vecteur u au point $x^\star$ que l'on note aussi par $D_u f(x^\star)$. On a donc:*

$$D_u f(x^\star) = \varphi'(0) = \lim_{t \to 0} \frac{\varphi(t) - \varphi(0)}{t} = \lim_{t \to 0} \frac{f(x^\star + tu) - f(x^\star)}{t}$$

Définition 3 *On appelle dérivée partielle de f par rapport à x_i au point $x^\star$, le vecteur de $\mathbb{R}^m$ défini par:*

$$\frac{\partial f}{\partial x_i}(x^\star) = D_{e_i} f(x^\star)$$

où e_i est le ième vecteur de la base canonique de $\mathbb{R}^m$ (vecteur colonne de $\mathbb{R}^m$ dont toutes les composantes sont nulles exceptée la ième). On a

$$\frac{\partial f}{\partial x_i}(x^\star) = \begin{pmatrix} \frac{\partial f_1}{\partial x_i}(x^\star) \\ \frac{\partial f_2}{\partial x_i}(x^\star) \\ \vdots \\ \frac{\partial f}{\partial x_i}(x^\star) \end{pmatrix}$$

• **Exemple d'une application non continue en un point et admettant toutes les dérivées directionnelles en ce point**

Soit $f : \mathbb{R}^2 \to \mathbb{R}$ définie par:

$$f(x_1, x_2) = \frac{x_1^6}{(x_2 - x_1^2)^2 + x_1^8} \text{ si } (x_1, x_2) \neq (0,0);\ f(0,0) = 0$$

Soit $u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \in \mathbb{R}^2$. Avec $x^\star = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

1) Montrer que f admet en $x^\star$ une dérivée selon tout vecteur u.

2) Montrer que f n'est pas continue en $x^\star$.
Indication: on pourra considérer $x_1(t) = t$ et $x_2(t) = t^2$.

Définition 4 *On dit que f est dérivable en x s'il existe une application linéaire de $R^n \to I\!R^m$ notée $f'(x)$ telle que:*

$$f(x+h) = f(x) + f'(x)h + \|h\|\alpha(h) \text{ avec } \lim_{h\to 0} \alpha(h) = 0$$

Lorsque f est derivable en tout point x appartenant à un ouvert Ω de $I\!R^n$, on dit que f est dérivable sur Ω.

Proposition 1 *Si $f : \Omega \to I\!R^m$ est dérivable en $x^\star$, alors:*
(a) f est continue en $x^\star$.
(b) f admet en $x^\star$ une dérivée selon tout vecteur $u \in I\!R^n$ et
$D_u f(x^\star) = f'(x^\star)u$
(c) la dérivée $f'(x^\star)$ est unique.

Proposition 2 *Soit Ω un ouvert de $\mathbb{R}^n$ et* $f = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix} : \Omega \to \mathbb{R}^m$ *où* $f_i : \Omega \to \mathbb{R}$ *pour* $i = 1, \ldots, m$. *On suppose que f est dérivable en $x^\star \in \Omega$. Alors la matrice associée à la dérivée $f'(x^\star)$ est la matrice rectangulaire $m \times n$ suivante dite matrice jacobienne.*

$$J_f(x^\star) = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(x^\star) & \dfrac{\partial f_1}{\partial x_2}(x^\star) & \ldots & \dfrac{\partial f_1}{\partial x_n}(x^\star) \\ \ldots & \ldots & \ldots & \ldots \\ \dfrac{\partial f_m}{\partial x_1}(x^\star) & \dfrac{\partial f_m}{\partial x_2}(x^\star) & \ldots & \dfrac{\partial f_m}{\partial x_n}(x^\star) \end{pmatrix}$$

Preuve: Soit $u = \sum_{i=1}^{n} u_i e_i$ un vecteur de $\mathbb{R}^n$. On a:

$$
\begin{aligned}
f'(x^\star)u &= f'(x^\star)(\sum_{i=1}^{n} u_i e_i) = \sum_{i=1}^{n} u_i f'(x^\star)e_i = \sum_{i=1}^{n} u_i D_{e_i} f(x^\star) \\
&= \sum_{i=1}^{n} u_i \frac{\partial f}{\partial x_i}(x^\star) \\
&= u_1 \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x^\star) \\ \vdots \\ \frac{\partial f_m}{\partial x_1}(x^\star) \end{pmatrix} + \ldots + u_n \begin{pmatrix} \frac{\partial f_1}{\partial x_n}(x^\star) \\ \vdots \\ \frac{\partial f_m}{\partial x_n}(x^\star) \end{pmatrix} \\
&= \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x^\star) & \frac{\partial f_1}{\partial x_2}(x^\star) & \ldots & \frac{\partial f_1}{\partial x_n}(x^\star) \\ \ldots & \ldots & \ldots & \ldots \\ \frac{\partial f_m}{\partial x_1}(x^\star) & \frac{\partial f_m}{\partial x_2}(x^\star) & \ldots & \frac{\partial f_m}{\partial x_n}(x^\star) \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}
\end{aligned}
$$

Nous allons donner maintenant une condition suffisante pour que f soit dérivable en $x^\star$.

Proposition 3 *Si, pour tout $i = 1, \ldots, n$ et tout $j = 1, \ldots, m$, on a:*
(a) les dérivées partielles $\dfrac{\partial f_j}{\partial x_i}(x^\star)$ sont définies sur une boule de centre $x^\star$,
(b) les application $x \mapsto \dfrac{\partial f_j}{\partial x_i}(x)$ sont continues en $x^\star$,
alors f est dérivable en $x^\star$.

Définition 5 *On dit que f est de classe $\mathcal{C}^1$ sur Ω si ses dérivées partielles existent et sont continues sur Ω.*

Applications de $\mathbb{R}^n$ dans $\mathbb{R}$:

• **Gradient d'une fonction:**

Soit Ω un ouvert de $\mathbb{R}^n$, f: $\Omega \to \mathbb{R}$ et $x^\star$ un point de Ω.

Définition 6 *On appelle vecteur gradient (ou tout simplement gradient) de f au point $x^\star$ que l'on note par $\nabla f(x^\star)$ le vecteur colonne suivant:*

$$\nabla f(x^\star) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x^\star) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x^\star) \end{pmatrix}$$

• Exprimer la matrice jacobienne de f en $x^\star$, $J_f(x^\star)$, à l'aide de son gradient en $x^\star$, $\nabla f(x^\star)$.

• **Matrice hessienne d'une fonction:**

Définition 7 *On dit que f est deux fois dérivable en $x^\star$ si l'application $T : x \to \nabla f(x)$ est dérivable en $x^\star$.*

On appelle matrice hessienne (ou tout simplement hessienne) de f en $x^\star$ que l'on note par $H_f(x^\star)$ la matrice jacobienne de T en $x^\star$:

$$H_f(x^\star) = \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(x^\star) & \dfrac{\partial^2 f}{\partial x_2 \partial x_1}(x^\star) & \dots & \dfrac{\partial^2 f}{\partial x_n \partial x_1}(x^\star) \\ \dots & \dots & \dots & \dots \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(x^\star) & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(x^\star) & \dots & \dfrac{\partial^2 f}{\partial x_n^2}(x^\star) \end{pmatrix}$$

On a donc par définition:

$$\nabla f(x^\star + h) - \nabla f(x^\star) = H_f(x^\star)h + \|h\|\alpha(h) \quad \text{avec} \quad \lim_{h \to 0} \alpha(h) = 0$$

Définition 8 *On appelle dérivée seconde de f en $x^\star$, notée $f''(x^\star)$, la forme bilinéaire de $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ associée à la matrice symétrique $H_f(x^\star)$.*

Proposition 4 (Théorème de Schwartz)

Si f est deux fois dérivable, alors:

- $f''(x^\star)(x, y) = f''(x^\star)(y, x)$ *ou encore*
- $\dfrac{\partial^2 f}{\partial x_i \partial x_j}(x^\star) = \dfrac{\partial^2 f}{\partial x_j \partial x_i}(x^\star)$ *pour tout* $i, j = 1, \ldots, n$.

Autrement dit, la matrice hessienne $H_f(x^\star)$ est symétrique.

Nous allons donner maintenant une condition suffisante pour que f soit deux fois dérivable en $x^\star$.

Proposition 5 *Si, pour tout $i = 1, \ldots, n$ et tout $j = 1, \ldots, n$, on a:*
(a) les dérivées partielles $\dfrac{\partial^2 f}{\partial x_i \partial x_j}(x^\star)$ sont définies sur une boule de centre $x^\star$,
(b) les applications $x \mapsto \dfrac{\partial^2 f}{\partial x_i \partial x_j}(x)$ sont continues en $x^\star$,
alors f est deux fois dérivable en $x^\star$.

Définition 9 *On dit que f est de classe $\mathcal{C}^2$ sur Ω si ses dérivées partielles secondes existent et sont continues sur Ω.*

Quelques théorèmes sur les applications dérivables:

Proposition 6 *Soient* f*:* $\Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$ *et* g*:* $\Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$ *deux applications dérivables en* $x^\star$*. Alors* $\forall(\alpha, \beta) \in \mathbb{R}^2$*,* $\alpha f + \beta g$ *est dérivable en* $x^\star$ *et on a :*

- $(\alpha f + \beta g)'(x^\star) = \alpha f'(x^\star) + \beta g'(x^\star)$
- $J_{\alpha f+\beta g}(x^\star) = \alpha J_f(x^\star) + \beta J_g(x^\star)$

Proposition 7 *Soient* $\Omega_1 \subseteq \mathbb{R}^n$ *et* $\Omega_2 \subseteq \mathbb{R}^m$ *deux ouverts,* f*:* $\Omega_1 \to \Omega_2$ *derivable en* $x^\star$ *et* g*:* $\Omega_2 \to \mathbb{R}^q$ *derivable en* $y^\star = f(x^\star) \in \Omega_2$*. Alors, l'application composée* $g \circ f$ *est dérivable en* $x^\star$ *et on a:*

- $(g \circ f)'(x^\star) = g'(y^\star) \circ f'(x^\star) = g'(f(x^\star)) \circ f'(x^\star)$
- $J_{g\circ f}(x^\star) = J_g(f(x^\star))J_f(x^\star)$

Soit Ω un ouvert de $\mathbb{R}^n$, x et u deux éléments de Ω et soit l'ensemble $I = \{t \in \mathbb{R} / x + tu \in \Omega\}$. On définit la fonction d'une variable scalaire suivante $\varphi : I \to \mathbb{R}^m$, $t \mapsto \varphi(t) = f(x + tu)$.

En introduisant la fonction affine $\alpha : \mathbb{R} \to \mathbb{R}^n$, $t \mapsto \alpha(t) = x + tu$, on voit que α est continue sur $\mathbb{R}$ et que $I = \alpha^{-1}(\Omega)$ qui est donc un ouvert de $\mathbb{R}$ contenant 0.

Théorème 1

(a) Si f est dérivable en $x + tu$, alors $\varphi(t) = f(x + tu)$ est dérivable en t et :

$$\varphi'(t) = \nabla f(x + tu)^T u$$

En particulier, φ est $\mathcal{C}^1$ sur I dès que f est $\mathcal{C}^1$ sur Ω

(b) si en outre f est $\mathcal{C}^2$ sur Ω, alors φ est $\mathcal{C}^1$ sur I et pour tout $t \in I$, on a:

$$\varphi''(t) = u^T \nabla^2 f(x + tu) u$$

Preuve:
(a) on a $\varphi(t) = (f \circ \alpha)(t)$. D'où $\varphi'(t) = f'(\alpha(t)) \circ \alpha'(t)$. Or, comme $\alpha' = u$ et $f'(\alpha(t)) = \nabla f(\alpha(t))^T$, on obtient:

$$\varphi'(t) = \nabla f(\alpha(t))^T u = \nabla f(x + tu)^T u$$

(b) Il suffit de remarquer que $\varphi'(t)$ s'écrit aussi comme $\varphi'(t) = (u^T \nabla f \circ \alpha)(t)$. On obtient alors:

$$\begin{aligned} \varphi''(t) &= (u^T \nabla f)'(\alpha(t)) \circ \alpha'(t) \\ &= u^T \nabla^2 f(\alpha(t)) u \\ &= u^T \nabla^2 f(x + tu) u \end{aligned}$$

Ceci termine la preuve du théorème.

L'utilisation de ce théorème ramène le calcul du gradient ou de la matrice hessienne d'une fonction à plusieurs variables au calcul de dérivées première et seconde d'une fonction à une seule variable. En effet, on a:

- $\varphi'(0) = \nabla f(x)^T u$
- $\varphi''(0) = u^T \nabla^2 f(x) u$

- Exemple: Calcul du gradient et de la hessienne de $f(x) = \|x\|_2 = \left(x^T x\right)^{1/2}$.

Considérons la fonction $\varphi(t) = f(x+tu) = (x^Tx + 2tx^Tu + t^2u^Tu)^{1/2}$.

- $\varphi'(t) = \dfrac{1}{2}\dfrac{2x^Tu + 2tu^Tu}{(x^Tx + 2tx^Tu + t^2u^Tu)^{1/2}} = \dfrac{x^Tu + tu^Tu}{(x^Tx + 2tx^Tu + t^2u^Tu)^{1/2}}$
- $\varphi''(t) = \dfrac{u^Tu}{(x^Tx + 2tx^Tu + t^2u^Tu)^{1/2}} - \dfrac{\left(x^Tu + tu^Tu\right)^2}{(x^Tx + 2tx^Tu + t^2u^Tu)^{3/2}}$
- $\varphi'(0) = \dfrac{x^Tu}{(x^Tx)^{1/2}} = \dfrac{x^Tu}{\|x\|}$
- $\varphi''(0) = \dfrac{u^Tu}{\|x\|} - \dfrac{(x^Tu)^2}{\|x\|^3} = \dfrac{u^Tu}{\|x\|} - \dfrac{(x^Tu)(x^Tu)}{\|x\|^3} =$

$u^T(\dfrac{1}{\|x\|}I_n - \dfrac{1}{\|x\|^3}xx^T)u$

où I_n est la matrice identité $n \times n$. On en déduit finalement:

- $\nabla f(x) = \dfrac{x}{\|x\|}$ et $\nabla^2 f(x) = \dfrac{1}{\|x\|}I_n - \dfrac{1}{\|x\|^3}xx^T$

Formules de Taylor-Lagrange:

Lemme 2 *Soit φ: $[0,1] \to \mathbb{R}$ une application dérivable telle φ' existe et est continue sur $[0,1]$ et φ'' existe sur $[0,1]$. Alors, il existe $\theta \in]0,1[$ tel que*

$$\varphi(1) = \varphi(0) + \varphi'(0) + \frac{1}{2}\varphi''(\theta)$$

Proposition 8 *: Formule de Taylor-Lagrange*
Soit f: $\Omega \subseteq \mathbb{R}^n \to \mathbb{R}$, $x^\star \in \Omega$ et h tel que $[x^\star, x^\star + h[\subset \Omega$. Si ∇f existe et est continu sur $[x^\star, x^\star + h[$ et si $H_f(x)$ existe pour tout $x \in]x^\star, x^\star + h[$, alors il existe $\theta \in]0,1[$, tel que:

$$f(x^\star + h) = f(x^\star) + \nabla f(x^\star)^T h + \frac{1}{2} h^T H_f(x^\star + \theta h) h$$

Preuve:

Soit $\varphi(t) = f(x^\star + th)$. Par hypothèse, φ' existe et est continue sur $[0,1]$ et φ'' existe sur $]0,1[$. D'après le lemme, il existe $\theta \in]0,1[$ tel que $\varphi(1) = \varphi(0) + \varphi'(0) + \frac{1}{2}\varphi''(\theta)$, d'où le résultat.

CHAPITRE 3

EXTREMA LOCAUX DE FONCTIONS DE PLUSIEURS VARIABLES

Rappels et Définitions

Définition 10 (matrice DP, DN)

Soit A une matrice carrée d'ordre n. La matrice A est dite Définie Positive, DP, (resp. Définie Négative, DN) si et seulement si:

$$\forall x \in I\!R^n \backslash \{0\} : x^T Ax > 0 \ (resp.\ x^T Ax < 0)$$

Définition 11 (matrice SDP, SDN)

Soit A une matrice carrée d'ordre n. La matrice A est dite Semi-Définie Positive, SDP, (resp. Semi-Définie Négative, SDN) si et seulement:

$$\forall x \in I\!R^n \backslash \{0\} : x^T Ax \geq 0 \ (resp.\ x^T Ax \leq 0)$$

Dans ce qui suit, soit $\mathcal{E} \subset \mathbb{R}^n$ un ensemble non vide et soit f une fonction de $\mathcal{E}$ dans $\mathbb{R}^q$.

Définition 12 (minimum, maximum, extremum d'une fonction)
On dit qu'un point $x^\star$ est un minimum (resp. maximum) de f, ou, de manière équivalente, que $x^\star$ minimise f (resp. maximise f) sur $\mathcal{E}$ si et seulement si: $\forall x \in \mathcal{E}$**:** $f(x^\star) \leq f(x)$ **(resp.** $\geq$**).**

On appelle extremum de f sur $\mathcal{E}$ un minimum ou un maximum de f.

Définition 13 (minimum, maximum local)
On dit qu'un point $x^\star$ est un minimum local (resp. maximum local) de f si et seulement si $x^\star$ est minimum de f sur une boule ouverte $\mathcal{B} \subset \mathcal{E}$.

Définition 14 (point critique)
On appelle point critique (ou stationnaire) de f un point $x^\star$ en lequel f est dérivable et $\nabla f(x^\star) = 0$

Conditions nécessaires d'optimalité locale

Soit Ω un ouvert de $\mathbb{R}^n$ et $x^\star \in \Omega$.

Théorème 2 (Principe de Fermat et condition du second ordre)

• *Si $x^\star$ est un extremum de f sur Ω et f est dérivable en $x^\star$, alors $x^\star$ est un point critique de f.*

• *Si en outre f est deux fois dérivable en $x^\star$, et $x^\star$ minimise (resp. maximise) f sur Ω, alors $\nabla^2 f(x^\star)$ est SDP (resp. SDN).*

La réciproque du principe de Fermat est fausse: un point critique de la fonction f n'est pas nécessairement un extremum local, même si la hessienne en ce point est SDP ou SDN. En effet, considérons la fonction $f(x_1, x_2) = x_1^3 + x_2^2$. Le point $(0, 0)$ est un point critique de f et le hessien de f en ce point est SDP. Pourtant, $(0, 0)$ n'est pas un extremum de f.

Conditions suffisantes d'optimalité locale

Soient x et y deux éléments de $\mathbb{R}^n$. On appelle segment d'extrémités x, y le sous-ensemble de $\mathbb{R}^n$ ainsi défini:

$$\{z \in \mathbb{R}^n / \exists t \in [0,1] : z = tx + (1-t)y\}$$

On dit que $[x, y]$ est le segment joignant x à y dans $\mathbb{R}^n$.

Définition 15 (Ensembles convexes)
Une partie de $\mathbb{R}^n$ est dite convexe si elle contient le segment joignant deux quelconques de ses points.

Théorème 3 *Soit Ω un ouvert convexe de $\mathbb{R}^n$.*
• *Si $x^\star$ est un point critique de f, f est $\mathcal{C}^2$ et $\nabla^2 f(\xi)$ est SDP (resp. SDN) en tout point ξ de Ω, alors $x^\star$ minimise (resp. maximise) f sur Ω.*
• *Si $\nabla^2 f(\xi)$ est DP (resp. DN) en tout point ξ de Ω, $x^\star$ est l'unique minimum (resp. maximum) de f dans Ω.*

• **Preuve:**

Posons $I = \{t \in \mathbb{R} / x^\star + tu \in \Omega\}$. Supposons $\nabla^2 f(\xi)$ SDP en tout point ξ de Ω. Pour tout $u \in \mathbb{R}$ et tout t dans I, posons
$\varphi(t) = f(x^\star + tu)$. On a:
$\varphi(t) = \varphi(0) + \frac{t^2}{2}\varphi''(\theta)$ où $\theta \in]0, t[$. Mais: $\varphi''(\theta) = u^T \nabla^2 f(x + \theta u) u$ est positif par hypothèse, donc:

$$\varphi(t) \geq \varphi(0) \quad \text{c'est-à-dire } f(x^\star + tu) \geq f(x^\star) \quad \text{dès que} \quad x^\star + tu \in \Omega$$

Si $\nabla^2 f(x^\star)$ est DP, l'inégalité stricte est vérifiée dès que: $x^\star + tu$ est distinct de $x^\star$, donc $x^\star$ est l'unique minimum de f sur Ω. On raisonne de manière équivalente lorsque $\nabla^2 f(x^\star)$ est DN ou SDN.

• **Exemple:**

Soit $f(x_1, x_2) = 2x_1 + 2x_2 + \dfrac{1}{x_1 x_2}$. On a

$H_f(x_1, x_2) = \begin{pmatrix} \dfrac{2}{x_2 x_1^3} & \dfrac{1}{x_2^2 x_1^2} \\ \dfrac{1}{x_2^2 x_1^2} & \dfrac{2}{x_1 x_2^3} \end{pmatrix}$ qui est DP en tout point de l'ouvert convexe $\Omega = \{(x_1, x_2) \in \mathbb{R}^2 / x_1, x_2 > 0\}$, et f admet un unique point critique $((1/2)^{1/3}, (1/2)^{1/3})$ dans Ω.

Points critiques non dégénérés

Définition 16 (point critique non dégénéré)
Un point critique $x^\star$ de f est dit non dégénéré si f est de classe $\mathcal{C}^2$ sur une boule ouverte de centre $x^\star$ et $\nabla^2 f(x^\star)$ est inversible.

Définition 17 (Extremum local strict)
On dit que $x^\star$ est un minimum ou maximum local strict de f si c'est l'unique minimum ou maximum local de f sur une boule ouverte de centre $x^\star$.

Théorème 4 : Conditions suffisantes d'optimalité locale
Si $x^\star$ est un point critique non dégénéré de f et $\nabla^2 f(x^\star)$ est DP (resp. DN), alors $x^\star$ est un minimum (resp. maximum) local strict de f.

Preuve du théorème:

Par continuité, il existe une boule de centre $x^\star$ sur laquelle $\nabla^2 f(x^\star)$ est DP (resp. DN). Toute boule dans $\mathbb{R}^n$ étant convexe, il suffit d'appliquer le théorème 3.

Exemple: le point $(0,0)$ est un minimum local strict de $f(x_1,x_2) = x_1^2 + x_2^2$ mais pas de $g(x_1,x_2) = (x_1+x_2)^2$.

CHAPITRE 4

MINIMISATION D'UNE FONCTION D'UNE VARIABLE

Directions de descente

Définition 18 *Soit $f : \mathbb{R}^n \to \mathbb{R}$. On dit qu'un vecteur $d \in \mathbb{R}^n$ est une direction de descente pour f au point $x^\star \in \mathbb{R}^n$ si:* $\exists \delta > 0; \forall \lambda \in]0, \delta[$: $f(x^\star + \lambda d) < f(x^\star)$.

Lorsque la fonction f est différentiable, on peut donner une condition simple qui garantit qu'une certaine direction est une direction de descente en un point $x^\star$.

Proposition 9 *Soit $f : \mathbb{R}^n \to \mathbb{R}$. Supposons que ∇f existe et est continu. Soient $x^\star, d \in \mathbb{R}^n$. Si $\nabla f(x^\star)^T d < 0$, alors le vecteur d est une direction de descente pour f en $x^\star$.*

Preuve: par définition, nous avons:

$$\begin{aligned} \lim_{\lambda \to 0^+} \frac{f(x^\star + \lambda d) - f(x^\star)}{\lambda} &= \nabla f(x^\star)^T d \\ &< 0 \quad \text{par hypothèse} \end{aligned}$$

D'après la définition précédente, on peut dire que:

• si $x^\star$ est un minimum local de f, alors il n'existe aucune direction de descente pour f en $x^\star$.

• S'il n'existe aucune direction de descente pour f au point $x^\star$ alors $x^\star$ est un point critique. Si en outre, $x^\star$ est non dégénéré, alors $x^\star$ est un minimum local qui de plus est strict. Si $x^\star$ est dégénéré, on ne peut pas conclure. En effet, la fonction $f(x_1, x_2) = 2x_1^4 - 3x_1^2x_2 + x_2^2 = (x_2 - 2x_1^2)(x_2 - x_1^2)$ n'admet aucune direction de descente en $(0, 0)$, et on peut vérifier que $(0, 0)$ n'est pas un minimum local de f.

Analyse de l'erreur

Considérons une suite vectorielle $\{x_k\}_{k\geq 0}$ dans $\mathbb{R}^n$ qui converge vers $x^\star$. Pour tout $k \geq 0$, on définit l'erreur $e_k = x_k - x^\star$.

Définition 19 vitesse et erreur asymptotique de convergence

S'il existe deux constantes $\alpha > 0$ *et* $\lambda \geq 0$ *telles que:* $\lim_{k\to 0} \dfrac{\|e_{k+1}\|}{\|e_k\|^\alpha} = \lambda$ *alors on dit que la suite* $\{x_k\}_{k\geq 0}$ *converge vers* $x^\star$ *avec une vitesse de convergence d'ordre* α *et une erreur asymptotique de convergence* λ.

En général, plus l'ordre de convergence est élevé, plus la vitesse de convergence est rapide. L'erreur asymptotique affecte aussi la vitesse de convergence mais de façon moins importante que l'ordre. Trois cas sont souvent rencontrés:

- $\alpha = 1$, la convergence est alors dite linéaire.
- $\alpha = 2$, la convergence est alors dite quadratique.
- $1 < \alpha < 2$, la convergence est alors dite superlinéaire.

Algorithmes de descente

Partant d'un point initial x_0, un algorithme de descente calcule un nouveau point x_1 de façon à réduire la valeur du critère à minimiser. Ce processus est répété à chaque étape en calculant le nouveau point x_{k+1} à partir de x_k. Il se résume comme suit:

- $k = 0$, $x_k \longleftarrow x_0$
- **Tant que** TESTARRET=FALSE **faire**

Calculer une direction de descente d_k au point x_k

Calculer un pas de descente α_k tel que $f(x_k + \alpha_k d_k) < f(x_k)$

$x_{k+1} \longleftarrow x_k + \alpha_k d_k$; mettre à jour TESTARRET; $x_k \longleftarrow x_{k+1}$

Fin Tant que

Un algorithme de descente est essentiellement déterminé par:

- la stratégie pour le choix des directions de descente successives.
- la stratégie pour le choix du pas dans la direction de descente.

Fonctions convexes et fonctions unimodales

Dans tout ce qui suit, I désigne un intervalle de $\mathbb{R}$ et f est une fonction de I dans $\mathbb{R}$.

Définition 20 (Fonction convexe)
La fonction f est dite convexe (resp. strictement convexe) sur I si et seulement si :

$$\forall \alpha \in]0,1[, \quad \forall s,t \in I, s \neq t, \quad f(\alpha s + (1-\alpha)t) \leq \alpha f(s) + (1-\alpha) f(t) \quad (resp. <)$$

Par analogie, on dit que f est concave (resp. strictement concave) sur I si et seulement si : $-f$ est convexe (resp. strictement convexe) sur I.

Définition 21 (Epigraphe)
On appelle épigraphe f l'ensemble $E_f = \{(x,y) \in I \times \mathbb{R} / f(x) \leq y\}$

Proposition 10 (propriétés des fonctions convexes)

- *Une fonction est convexe si et seulement si son épigraphe est convexe.*
- *Si f est convexe sur I alors elle est continue sur I.*
- *Si f est dérivable sur I, alors elle y est $\mathcal{C}^1$.*
- *Soit $f \in \mathcal{C}^1(I)$. Alors f est convexe (resp. strictement convexe) sur I si et seulement si f' est croissante (resp. strictement croissante) sur I.*
- *La somme de deux fonctions convexes sur I est convexe sur I.*
- *Le produit de deux fonctions convexes sur I n'est pas nécessairement convexe sur I (on peut le vérifier avec $t \mapsto t$ et $t \mapsto t^2$ sur $]-1,1[$).*

Théorème 5 *Si f est convexe sur I, et admet un point critique $t^\star$ dans l'intérieur de I, alors $t^\star$ minimise f sur I.*

Preuve du théorème 5:

Soit $\alpha \in]0,1[$ et $t, s \in I$. Comme f est convexe sur I, on :

$$f(\alpha s + (1-\alpha)t^\star) \leq \alpha f(s) + (1-\alpha) f(t^\star)$$

mais comme $f(\alpha s + (1-\alpha)t^\star) = f(t^\star + \alpha(s - t^\star))$, on obtient

$$f(t^\star + \alpha(s-t^\star)) - f(t^\star) \leq \alpha(f(s) - f(t^\star))$$

d'où

$$\lim_{\alpha \to 0} \frac{f(t^\star + \alpha(s-t^\star)) - f(t^\star)}{\alpha} \leq f(s) - f(t^\star)$$

Comme $t^\star$ est un point critique de f, la limite intervenant dans le membre gauche de l'inégalité précédente existe et elle est nulle et on obtient ainsi $f(t^\star) \leq f(s)$.

Définition 22 (Fonction unimodale) *La fonction f est dite unimodale sur I si et seulement si les deux propriétés suivantes sont vérifiées:*

- *f admet un unique minimum $t^\star$ dans l'intérieur de I.*
- *f est strictement décroissante sur $I \cup]-\infty, t^\star[$ et est strictement croissante sur $I \cup]t^\star, +\infty[$.*

Théorème 6 *Si f est strictement convexe sur I et atteint son minimum sur I en un point $t^\star$ dans l'intérieur de I alors f est unimodale sur I.*

Preuve: Il suffit de montrer la deuxième propriété de la définition 22. En effet, soit $s, t \in I \cup]-\infty, t^\star[$. Supposons que $s < t$ et montrons alors que $f(s) > f(t)$. Comme $t \in]s, t^\star[$, alors il existe $\alpha \in]0, 1[$ tel que $t = \alpha s + (1-\alpha)t^\star$. Or, comme f est strictement convexe sur I, on a:

$$f(t) < \alpha f(s) + (1-\alpha) f(t^\star) \leq f(s)$$

d'où $f(t) < f(s)$ et f est strictement décroissante sur $I \cup]-\infty, t^\star[$. On montre de la même façon que f est strictement croissante sur $I \cup]t^\star, +\infty[$.

Algorithme de la section dorée

Il s'agit de partir de trois points a, b, c tels que $f(a) > f(b) < f(c)$. On sait alors que l'on a un minimum dans l'intervalle $]a, c[$. On peut rétrécir cet intervalle en considérant un point x entre b et c. Mais deux choix sont alors possibles pour ce nouvel intervalle: $[a, x]$ ou $[b, c]$ comme illustré sur la figure suivante :

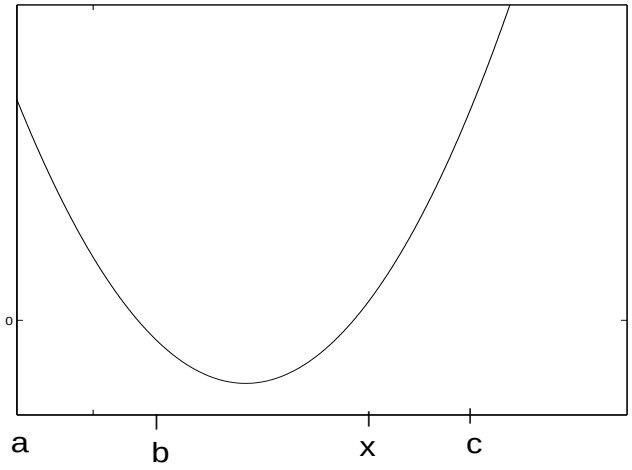


Figure 1: deux choix possibles pour le nouvel intervalle

Pour le choix du nouvel intervalle, on procède comme suit. Tout d'abord, l'intervalle $[a, c]$ est normalisé en se ramenant à l'intervalle $[0, 1]$ où les abscisses 0 et 1 sont respectivement associées aux extrémités de l'ancien intervalle a et c. Notons par w l'abscisse du point b et par $w + z$ celle du point x:

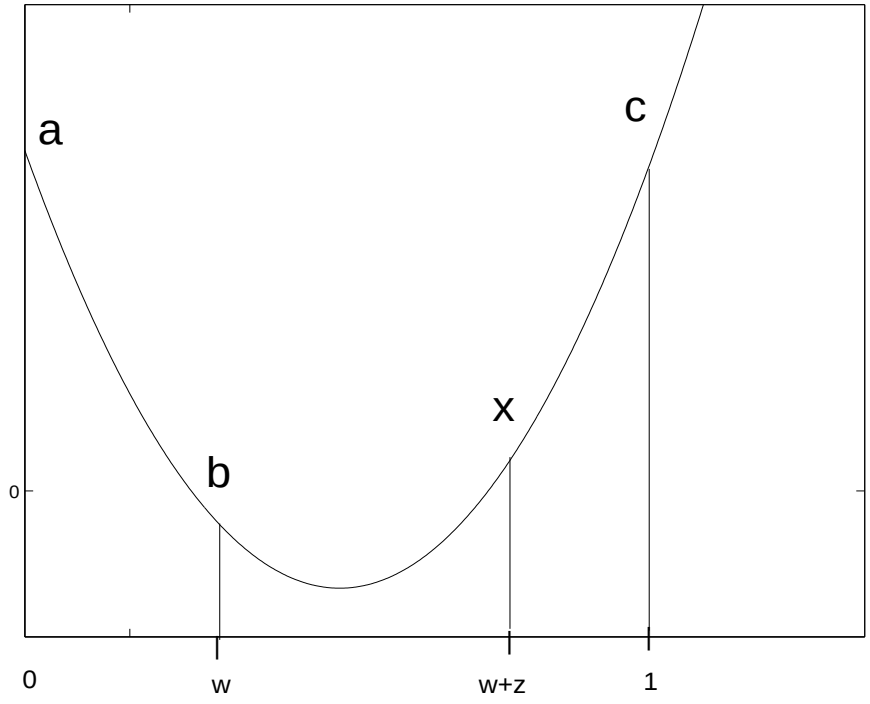
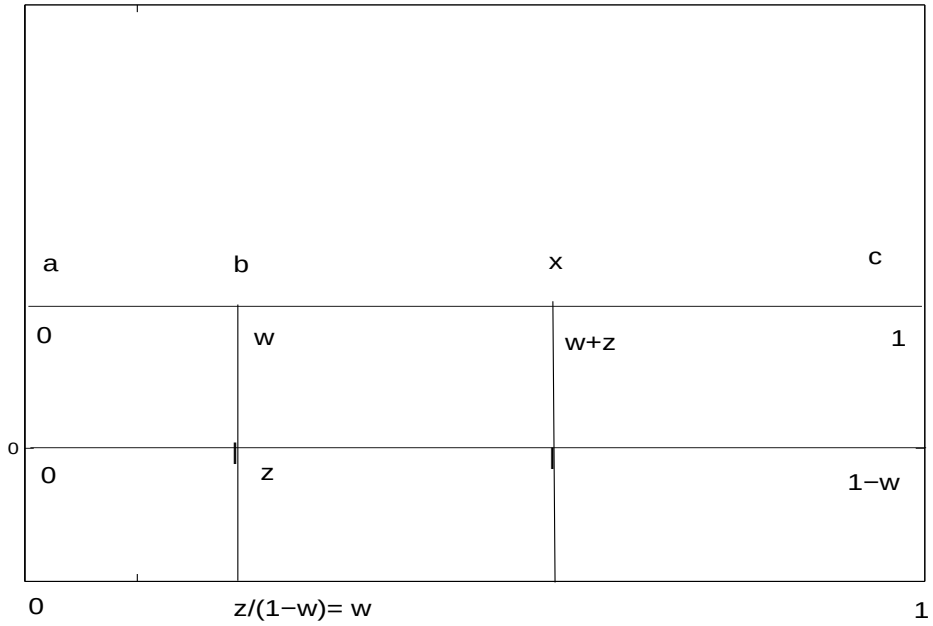


Figure 2: Normalisation des intervalles de recherche

Le problème consiste à choisir z pour minimiser la longueur du nouvel intervalle de recherche. Les deux longueurs possibles sont $(w+z)$ correspondant aux choix de l'intervalle $[a,x]$ et $(1-w)$ correspondant au choix de l'intervalle $[b,c]$. Le choix optimum est obtenu en cherchant le réel z tel que les deux longueurs possibles soient égales, c'est-à-dire $w+z=1-w$.



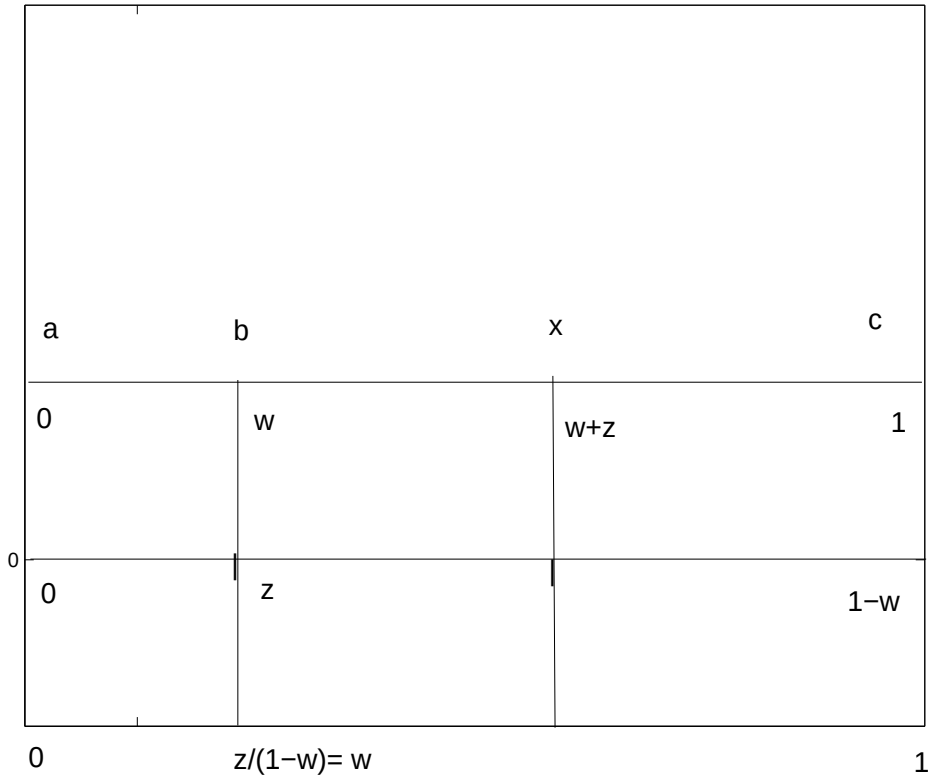
Mais d'un autre côté, le point b a été choisi à l'étape précédente selon le même processus, d'où après normalisation, on obtient: $w = \dfrac{z}{1-w} = \dfrac{1-2w}{1-w}$

La variable w satisfait donc à l'équation du second degré suivante:
$w^2 - 3w + 1 = 0$. La seule racine appartenant à l'intervalle $[0, 1]$ est
$w = \dfrac{3 - \sqrt{5}}{2} \simeq 0.38197$. En posant $\alpha = 1 - w \simeq 0.618$, on remarque que la longueur du nouvel intervalle de recherche est égale à $1 - w = w + z = \alpha$.
L'inverse de α est égal à $\dfrac{1}{\alpha} = \dfrac{1 + \sqrt{5}}{2}$ qui est le nombre d'or.

L'algorithme de la section dorée peut se se décrire comme suit. On part d'un intervalle $[a, b]$. On calcule deux nouvelles abscisses, c et x à l'intérieur de cet intervalle de sorte que l'une de ces abscisses soit l'extrémité du nouvel intervalle et de sorte que la longueur du nouvel intervalle soit égale à $\alpha(c - a)$ (réduction d'ordre α de l'intervalle de recherche). L'abscisse retenue parmi b et x sera celle pour laquelle le critère est le plus petit.

Les abscisses b et x sont calculées comme suit:

$$\begin{aligned} b &= a + w(c-a) \\ &= a + (1-\alpha)(c-a) \\ &= \alpha a + (1-\alpha)c \\ &= c - \alpha(c-a) \end{aligned} \qquad \begin{aligned} x &= a + (w+z)(c-a) \\ &= a + (1-w)(c-a) \\ &= a + \alpha(c-a) \\ &= a + c - b \end{aligned}$$



En appelant TOL la précision avec laquelle on souhaite avoir la solution et f le critère, l'algorithme peut se résumer donc comme suit:

DEBUT ALGO

$b = c - \alpha(c - a)$;
$x = a + c - b$;
Tant que $(c - a)/2 > TOL$ **faire**
Si: $f(b) < f(x)$ **alors** $c = x$; $x = b$; $b = a + c - x$; (élimination de c: [a, x] est le nouvel intervalle de recherche)
Sinon: $a = b$; $b = x$; $x = a + c - b$; (élimination de a: [b, c] est le nouvel intervalle de recherche)
Fin Tant que
Retourner: $(a + c)/2$.

FIN ALGO

Analyse de l'algorithme de la section dorée

L'algorithme de la section dorée permet de réduire à chaque itération la longueur de l'intervalle de recherche par un facteur égal à α. On peut se demander jusqu'à quelle précision pourrait-on avoir la solution désirée $x^\star$ qui réalise le minimum du critère. Pour ce faire, nous allons procéder à une analyse au voisinage de $x^\star$. En notant par f le critère, nous avons:

$$f(x) = f(x^\star) + \frac{(x - x^\star)^2}{2} f''(x^\star)$$

Le minimum $x^\star$ sera atteint avec une précision ε si l'on a:

$$\left| \frac{(x - x^\star)^2}{2} f''(x^\star) \right| \leq \varepsilon |f(x^\star)|$$

$$(x - x^\star)^2 \leq 2\varepsilon \frac{|f(x^\star)|}{|f''(x^\star)|} = \varepsilon |x^\star|^2 \frac{2|f(x^\star)|}{|x^\star|^2 |f''(x^\star)|}$$

ou de manière équivalente

$$|x - x^\star| \leq \sqrt{\varepsilon} |x^\star| \sqrt{\frac{2|f(x^\star)|}{|x^\star|^2 |f''(x^\star)|}}$$

Si l'on suppose que la quantité $\sqrt{\frac{2|f(x^\star)|}{|x^\star|^2 |f''(x^\star)|}}$ est de l'ordre de l'unité, on aura:

$$|x - x^\star| \leq \sqrt{\varepsilon} |x^\star|$$

En d'autres termes, $\sqrt{\varepsilon}$ est de l'ordre de la précision relative avec laquelle on souhaite obtenir la solution $x^\star$. Ceci nous invite à faire attention lors du choix de ε. Une valeur dont la racine carrée est plus petite que la précision de machine a peu d'intérêt puisqu'une telle précision ne peut pas être atteinte.

Remarques 1

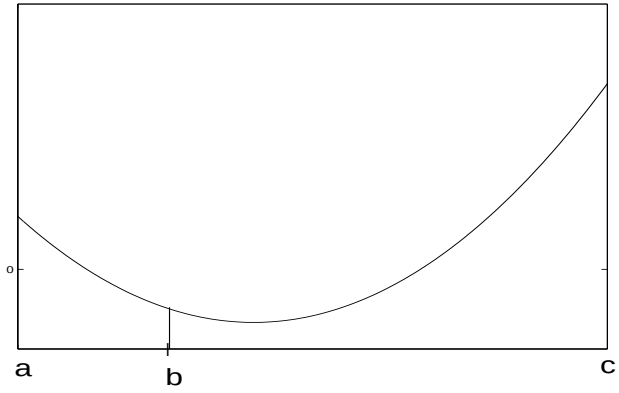
1) L'algorithme de la section dorée commence avec trois points $a < b < c$ tels que $f(a) > f(b) < f(c)$.Même si ces trois points de départ ne respectent pas pas le rapport $\alpha = \dfrac{|b-c|}{|c-a|}$, on a convergence vers ce rapport après l'application de l'algorithme.

2) La convergence de l'algorithme de la section dorée est linéaire.

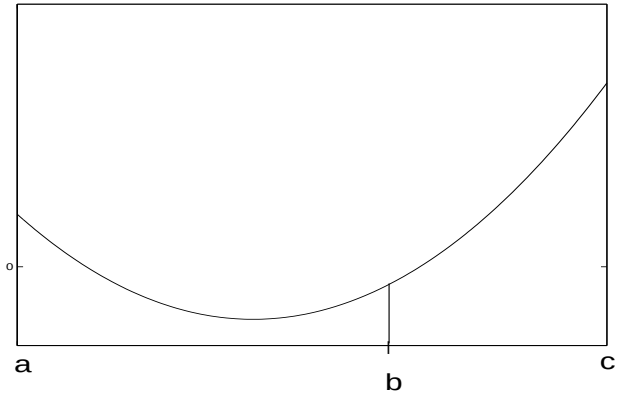
Théorème 7 *Si f est une fonction unimodale sur l'intervalle $[a, b]$, alors l'algorithme de la section dorée converge vers le minimum de f sur cet intervalle.*

Utilisation de la dérivée première

A première vue, on est tenté de ramener le problème de minimisation d'une fonction f à un problème de recherche de racine (de f'). Mais dans ce cas, il n'y a que la condition du premier ordre qui est satisfaite et on ne sait pas en particulier si l'on a affaire à un minimum ou maximum.



$f'(b) < 0 \Longrightarrow$ choisir $[b, c]$



$f'(b) > 0 \Longrightarrow$ choisir $[a, b]$

Si l'on suppose que l'on dispose d'un intervalle $[a, c]$ et du signe de la dérivée du critère à minimiser en $b \in]a, c[$, on peut à partir de cette information réduire l'intervalle de recherche en remplaçant l'une des extrémités de $[a, b]$ par c

CHAPITRE 5

ALGORITHME DU GRADIENT

Généralités

Soit $f : \mathbb{R}^n \to \mathbb{R}$ et soit $x^0 \in \mathbb{R}^n$ / $\nabla f(x^0) \neq 0$.

L'objectif consiste à chercher $\min_{x \in \mathbb{R}^n} f(x)$.

Soit $x \in \mathbb{R}^n$; nous avons:

$$\begin{aligned} f(x) &= f(x^0) + \nabla f(x^0)^T (x - x^0) + \mathcal{O}(\|x - x^0\|^2) \\ &\simeq f(x^0) + \nabla f(x^0)^T (x - x^0) \end{aligned}$$

dans un voisinage de x^0.

Considérons maintenant la droite, $x(\alpha)$ passant par x^0 et de pente $-\nabla f(x^0)$.

L'équation de cette droite est:

$$x(\alpha) = x^0 - \alpha \nabla f(x^0)$$

Evaluons $f(x)$ le long de cette droite. Nous avons:

$$\begin{aligned} f(x(\alpha)) &= f(x^0) + \nabla f(x^0)^T(-\alpha \nabla f(x^0)) \\ &= f(x^0) - \alpha \|\nabla f(x^0)\|^2 \\ &< f(x^0) \end{aligned}$$

pour $\alpha > 0$ suffisamment petit.

Donc la direction du gradient, avec un sens opposé au gradient, est une direction de descente. C'est même la meilleure localement (c-à-d pour $\alpha > 0$ suffisamment petit).

Algorithme du gradient à pas optimal

Soit x^0 un point initial. On pose:

$$x^1 = x^0 - \alpha_0 \nabla f(x^0) \quad \text{avec} \quad \alpha_0 = arg \min_{\alpha>0} f\left(x^0 - \alpha \nabla f(x^0)\right)$$

Le point x^1 ainsi obtenu est le meilleur possible dans la direction $\nabla f(x^0)$, d'où l'appellation de la "l'algorithme de la plus grande pente" ou du "gradient à pas optimal".

L'algorithme du gradient à pas optimal est donc un processus itératif qui consiste à construire une suite $\{x^k\}_{k\geq 0}$:

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) \quad \text{avec} \quad \alpha_k = arg \min_{\alpha>0} f\left(x^k - \alpha \nabla f(x^k)\right)$$

Propriétés de l'algorithme à pas optimal

$\mathcal{P}1$. Deux directions de descente successives sont orthogonales.

$\mathcal{P}2$. Dans le cas où le critère est quadratique, c'est-à-dire:

$$f(x) = \frac{1}{2}x^T Ax - b^T x \quad avec \quad A = A^T > 0$$

nous avons:

$$\alpha_k = \frac{\|\nabla f(x^k)\|^2}{\|\nabla f(x^k)\|_A^2} = \frac{\|Ax^k - b\|^2}{\|Ax^k - b\|_A^2} = \frac{\|Ax^k - b\|^2}{(Ax^k - b)^T A (Ax^k - b)}$$

• **Preuve**:

$\mathcal{P}1$. A l'étape $k+1$, pour obtenir le pas de descente optimal, l'algorithme minimise la fonction

$$\varphi(\alpha) = f(x^k + \alpha u^k) \quad \text{avec} \quad u^k = -\nabla f(x^k)$$

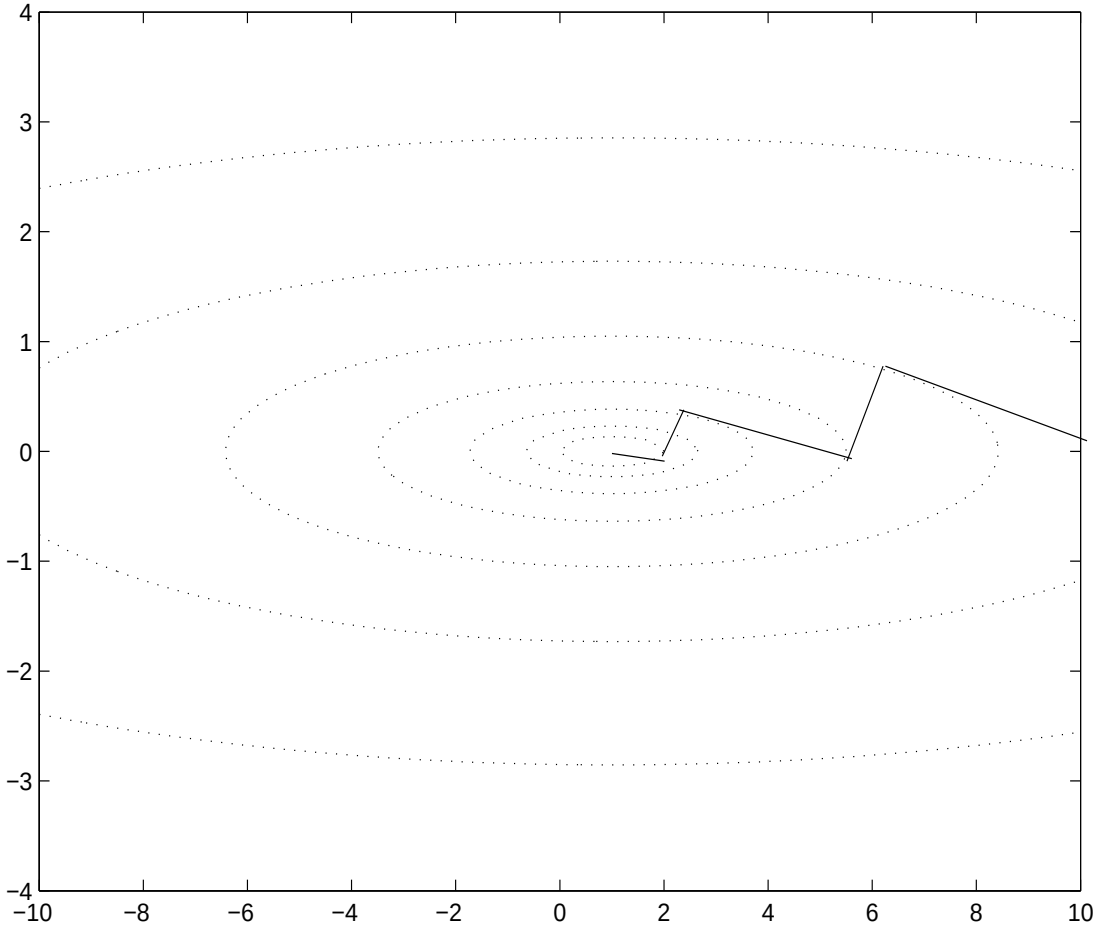
Si α_k est la pas optimal calculé, on a $\varphi'(\alpha_k) = 0$.

Mais, d'un autre cté, on a:

$$\begin{aligned} \varphi'(\alpha_k) &= \nabla f(x^k + \alpha_k u^k)^T u^k \\ &= -\nabla f(x^k - \alpha_k \nabla f(x^k))^T \nabla f(x^k) \\ &= -\nabla f(x^{k+1})^T \nabla f(x^k) \end{aligned}$$

d'où le résultat.

Cette propriété se traduit par des des évolutions (d'une itérée à l'autre) sous formes de zigzags lorsque l'on a affaire à des hessiennes mal conditionnées (lignes de niveau très aplaties).



$\mathcal{P}2$. Notons tout d'abord, que dans le cas quadratique, nous avons:
$\nabla f(x^k) = Ax^k - b$.
D'après la première propriété, nous avons $\nabla f(x^{k+1})^T \nabla f(x^k) = 0$. Développons le membre gauche de cette égalité:

$$\begin{aligned}
\nabla f(x^{k+1})^T \nabla f(x^k) &= (Ax^{k+1} - b)^T \nabla f(x^k) \\
&= \left(A\left(x^k - \alpha_k \nabla f(x^k)\right) - b\right)^T \nabla f(x^k) \\
&= \left(Ax^k - b - \alpha_k A \nabla f(x^k)\right)^T \nabla f(x^k) \\
&= \left(\nabla f(x^k) - \alpha_k A \nabla f(x^k)\right)^T \nabla f(x^k) \\
&= \|\nabla f(x^k)\|^2 - \alpha_k \left(\nabla f(x^k)\right)^T A \nabla f(x^k) \\
&= \|\nabla f(x^k)\|^2 - \alpha_k \|\nabla f(x^k)\|_A^2 \\
&= 0 \qquad \Longrightarrow
\end{aligned}$$

$$\alpha_k = \frac{\|\nabla f(x^k)\|^2}{\|\nabla f(x^k)\|_A^2}$$

Théorème (de l'algorithme à pas optimal)

Dans le cas quadratique :

$$f(x) = \frac{1}{2}x^T Ax - b^T x \quad \text{avec} \quad A = A^T > 0$$

l'algorithme du gradient à pas optimal converge vers l'optimum $x^\star = A^{-1}b$.

De plus, nous avons:

$$\|x^{k+1} - x^\star\|_A^2 \leq \|x^k - x^\star\|_A^2 \left(\frac{r-1}{r+1}\right)^2$$

où $r = \dfrac{\lambda_{max}(A)}{\lambda_{min}(A)}$ est le nombre de condionnement de A selon la norme euclidienne.

Pour la démonstration de ce théorème, nous avons besoin du lemme suivant:

Lemme 3 { Inégalité de Kantorovitch }
Si A est une matrice symétrique définie positive d'ordre n, de valeurs propres λ_i (positives) ordonnées dans le sens suivant $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, alors:

$$\forall u \in I\!R^n : \frac{\|u\|^4}{\|u\|_A^2 \|u\|_{A^{-1}}^2} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2} \qquad (\star)$$

Preuve:
Nous allons nous placer dans une base où A est diagonale. Soit $x \in \mathbb{R}^n$ tel que $\|x\| = 1$. Démontrons $(\star)$ pour x. Nous avons:

$$\frac{1}{\sum_{i=1}^n \lambda_i x_i^2 \sum_{i=1}^n x_i^2/\lambda_i} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}$$

Chaque terme des deux membres étant positif, on peut passer à leur racine:

$$\frac{1}{\sqrt{\sum_{i=1}^n \lambda_i x_i^2}\sqrt{\sum_{i=1}^n x_i^2/\lambda_i}} \geq \frac{2\sqrt{\lambda_1\lambda_n}}{(\lambda_1+\lambda_n)}$$

où encore
$$\sqrt{\sum_{i=1}^{n}\lambda_i x_i^2}\sqrt{\sum_{i=1}^{n}x_i^2/\lambda_i}\leq\frac{(\lambda_1+\lambda_n)}{2\sqrt{\lambda_1\lambda_n}}$$

ou de manière équivalente
$$\sqrt{\sum_{i=1}^{n}\lambda_i x_i^2}\sqrt{\sum_{i=1}^{n}\frac{\lambda_1\lambda_n}{\lambda_i}x_i^2}\leq\frac{(\lambda_1+\lambda_n)}{2}$$

Or, nous avons
$$\begin{aligned}\sqrt{\sum_{i=1}^{n}\lambda_i x_i^2}\sqrt{\sum_{i=1}^{n}\frac{\lambda_1\lambda_n}{\lambda_i}x_i^2} &\leq \frac{1}{2}\left(\left(\sqrt{\sum_{i=1}^{n}\lambda_i x_i^2}\right)^2+\left(\sqrt{\sum_{i=1}^{n}\frac{\lambda_1\lambda_n}{\lambda_i}x_i^2}\right)^2\right)\\ &= \frac{1}{2}\sum_{i=1}^{n}\left(\lambda_i+\frac{\lambda_1\lambda_n}{\lambda_i}\right)x_i^2\end{aligned}$$

Il suffit donc de montrer que $\frac{1}{2}\sum_{i=1}^{n}\left(\lambda_i+\frac{\lambda_1\lambda_n}{\lambda_i}\right)x_i^2-\frac{(\lambda_1+\lambda_n)}{2}\leq 0$

En effet, on a:

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^{n}\left(\lambda_i+\frac{\lambda_1\lambda_n}{\lambda_i}\right)x_i^2-\frac{(\lambda_1+\lambda_n)}{2} &= \frac{1}{2}\sum_{i=1}^{n}\left(\lambda_i+\frac{\lambda_1\lambda_n}{\lambda_i}\right)x_i^2-\frac{(\lambda_1+\lambda_n)}{2}\sum_{i=1}^{n}x_i^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}\left(\lambda_i+\frac{\lambda_1\lambda_n}{\lambda_i}-(\lambda_1+\lambda_n)\right)x_i^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}\frac{(\lambda_i^2+\lambda_1\lambda_n)-\lambda_i(\lambda_1+\lambda_n)}{\lambda_i}x_i^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}\frac{(\lambda_i-\lambda_1)(\lambda_i-\lambda_n)}{\lambda_i}x_i^2 \\
&\leq 0 \qquad \text{puisque} \quad \lambda_1\leq\lambda_i\leq\lambda_n
\end{aligned}$$

L'inégalité $(\star)$ est donc montré pour tous les x dont la norme est égale à 1.

Soit maintenant y quelconque dans $\mathbb{R}^n$ et montrons cette inégalité pour y.

Posons $x = \dfrac{y}{\|y\|}$. Nous avons:

$$\begin{aligned}
\frac{\|y\|^4}{\|y\|_A^2 \|y\|_{A^{-1}}^2} &= \frac{\|\|y\|x\|^4}{\|y\|^4 \|x\|_A^2 \|x\|_{A^{-1}}^2} \\
&= \frac{\|x\|^4}{\|x\|_A^2 \|x\|_{A^{-1}}^2} \\
&\geq \frac{4\lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2} \quad \text{puisque} \quad \|x\| = 1
\end{aligned}$$

Ceci termine la preuve du lemme.

Preuve du théorème

$$
\begin{aligned}
\|x^{k+1} - x^\star\|_A^2 &= \|x^k - \alpha_k \nabla f(x^k) - x^\star\|_A^2 \\
&= \|x^k - x^\star\|_A^2 + \|\alpha_k \nabla f(x^k)\|_A^2 - 2\alpha_k (x^k - x^\star)^T A \nabla f(x^k) \\
&= \|x^k - x^\star\|_A^2 + \alpha_k^2 \|\nabla f(x^k)\|_A^2 - 2\alpha_k (Ax^k - b)^T \nabla f(x^k) \\
&= \|x^k - x^\star\|_A^2 + \alpha_k^2 \|\nabla f(x^k)\|_A^2 - 2\alpha_k \|\nabla f(x^k)\|^2 \\
&= \|x^k - x^\star\|_A^2 + \left(\frac{\|\nabla f(x^k)\|^2}{\|\nabla f(x^k)\|_A^2} \right)^2 \|\nabla f(x^k)\|_A^2 \\
&\quad -2 \left(\frac{\|\nabla f(x^k)\|^2}{\|\nabla f(x^k)\|_A^2} \right) \|\nabla f(x^k)\|^2 \\
&= \|x^k - x^\star\|_A^2 + \frac{\|\nabla f(x^k)\|^4}{\|\nabla f(x^k)\|_A^2} - 2 \frac{\|\nabla f(x^k)\|^4}{\|\nabla f(x^k)\|_A^2} \\
&= \|x^k - x^\star\|_A^2 - \frac{\|\nabla f(x^k)\|^4}{\|\nabla f(x^k)\|_A^2}
\end{aligned}
$$

D'autre part, nous avons:

$$\begin{aligned}
\|x^k - x^\star\|_A^2 &= \left(x^k - x^\star\right)^T A \left(x^k - x^\star\right) \\
&= \left(Ax^k - Ax^\star\right)^T A^{-1} \left(Ax^k - Ax^\star\right) \\
&= \left(Ax^k - b\right)^T A^{-1} \left(Ax^k - b\right) \\
&= \left(\nabla f(x^k)\right)^T A^{-1} \left(\nabla f(x^k)\right) = \|\nabla f(x^k)\|_{A^{-1}}^2
\end{aligned}$$

On en déduit:

$$\begin{aligned}
\|x^{k+1} - x^\star\|_A^2 &= \|x^k - x^\star\|_A^2 \left(1 - \frac{\|\nabla f(x^k)\|^4}{\|\nabla f(x^k)\|_A^2 \|\nabla f(x^k)\|_{A^{-1}}^2}\right) \\
&\leq \|x^k - x^\star\|_A^2 (1 - \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2}) \quad \text{d'après l'inégalité de Kantorovitch} \\
&= \|x^k - x^\star\|_A^2 \frac{(\lambda_n - \lambda_1)^2}{(\lambda_1 + \lambda_n)^2} = \|x^k - x^\star\|_A^2 \left(\frac{r-1}{r+1}\right)^2
\end{aligned}$$

avec $r = \dfrac{\lambda_n}{\lambda_1}$. Ceci termine la preuve du théorème.

Le théorème du gradient optimal nous fournit un ordre du rapport de l'erreur $x^k - x^\star$ entre deux itérées successives. Cet ordre est égal à $\left(\frac{r-1}{r+1}\right)^2$ où r (≥ 1) est le nombre de condionnement de A par rapport à la norme euclidienne.

La fonction $r \mapsto \frac{r-1}{r+1}$ est strictement croissante entre 1 et $+\infty$ et elle prend la valeur 0 en 1 et la valeur 1 en $+\infty$. L'algorithme du gradient à pas optimal converge donc très lentement lorsque la matrice A est mal conditionnée (r est grand). En revanche, si $r = 1$ (en dimension 2, les lignes de niveau sont des cercles), la convergence se fait en un pas quelque soit le point initial.

Théorème 8 *(* **convergence de l'algorithme***)*
Si $\Omega \subset \mathbb{R}^n$ est un bassin d'ellipticité de f, l'algorithme du gradient à pas optimal converge, pour toute initialisation x_0 dans Ω, vers l'unique minimum local $x^\star$ de f sur Ω.

Algorithme du gradient à pas fixe

Il s'agit de se passer de la recherche du pas optimal, α_k, à chaque étape et de prendre un pas constant α pour toutes les itérations:

$$x^{k+1} = x^k - \alpha \nabla f(x^k), \quad \alpha > 0$$

Le choix de α se fait naturellement en fonction du problème considéré:

- si α est choisi "trop grand", l'algorithme risque de diverger,

- si α est choisi "trop petit", la convergence serait très lente.

Lorsque l'on dispose d'informations supplémentaires sur le critère à minimiser f (constante de Lipschitz de la dérivée, constante de forte convexité, etc.), on peut évaluer à l'avance le meilleur paramètre α possible.

Variantes de l'algorithme du gradient

Lorsque l'évaluation du critère en un point est compliquée, la procédure de recherche linéaire utilisée par l'algorithme du gradient à pas optimal peut se révéler coûteuse. C'est pourquoi, on préfère parfois utiliser des procédures de recherche linéaire qui permettent simplement de faire décroître suffisamment la valeur du critère. On présentera deux variantes:

- Recherche linéaire avec le critère d'Armijo

- Recherche linéaire avec l'interpolation parabolique

Recherche linéaire avec le critère d'Armijo

On se donne un réel ε compris entre 0 et 1, et on cherche un pas α vérifiant:

$$f(x+\alpha u) < f(x) + \varepsilon\alpha \left(\nabla f(x)\right)^T u$$

Théorème 9 critère d'Armijo
Si $\left(\nabla f(x)\right)^T u <$, *alors il existe* $\alpha^\star$ *tel que pour tout* $\alpha \in]0, \alpha^\star[$, *on a:*
$f(x+\alpha u) < f(x) + \varepsilon\alpha \left(\nabla f(x)\right)^T u$

Preuve: posons $\varphi(\alpha) = f(x+\alpha u) - f(x) - \varepsilon\alpha \left(\nabla f(x)\right)^T u$. Nous avons:
$\varphi'(\alpha) = \left(\nabla f(x+\alpha u)\right)^T u - \varepsilon \left(\nabla f(x)\right)^T u$. D'où: $\varphi'(0) = (1-\varepsilon) \left(\nabla f(x)\right)^T u < 0$.

Comme $\varphi(0) = 0$, alors pour $\alpha > 0$ suffisamment petit, on a $\varphi(\alpha) < 0$.

Algorithme

Pour l'algorithme de recherche linéaire utilisant le critère d'Aramijo, on peut se donner la séquence suivante pour les valeurs de α, $\alpha_i = \alpha_0 2^{-i}$ avec $\alpha_0 = 1$ et on choisit ensuite le plus grand α_i satisfaisant au critère. Cela donne l'algorithme suivant:

$\varepsilon = 0.5$; $\alpha = 1$; x^k, d^k donnés;

TANT QUE $f(x^k + \alpha d^k) > f(x^k) + \varepsilon\alpha \left(\nabla f(x^k)\right)^T d^k$ FAIRE
$\alpha \leftarrow \alpha/2$;
FIN TANT QUE

Recherche linéaire avec l'interpolation parabolique

L'idée générale consiste à chercher à chaque étape, le α qui réalise le minimum de la parabole passant par les trois points: $(\alpha_1, f(x^k + \alpha_1 d^k))$, $(\alpha_2, f(x^k + \alpha_2 d^k))$ et $(\alpha_3, f(x^k + \alpha_3 d^k))$:

- Le réel α_1 est pris égal à 0,
- Le réel α_3 est initialisé à 1 puis éventulellement divisé par 2 après chaque test jusqu'à ce que l'on ait $f(x^k + \alpha_3 d^k) < f(x^k + \alpha_1 d^k)$,
- Le réel α_2 est pris égal à $\alpha_3/2$.

Algorithme

$\alpha_1 = 0$; $\alpha_3 = 1$; $f_1 = f(x^k)$; $f_3 = f(x^k + \alpha_3 d^k)$;
TANT QUE $f_3 \geq f_1$ FAIRE
$\alpha_3 \leftarrow \alpha_3/2$; $f_3 \rightarrow f(x^k + \alpha_3 d^k)$;
FIN TANT QUE
$\alpha_2 = \alpha_3/2$; $f_2 = f(x^k + \alpha_2 d^k)$;
// Soit $P(\alpha)$ le polynme qui interpole $f(x^k + \alpha d^k)$ en α_1, α_2 et α_3;
// L'équation de $P(\alpha)$ est: $P(\alpha) = f_1 + h_1\alpha + h_3\alpha(\alpha - \alpha_2)$;

$h_1 = (f_2 - f_1)/\alpha_2$; $h_2 = (f_3 - f_2)/\alpha_3$; $h_3 = (h_2 - h_1)/(\alpha_3 - \alpha_2)$;
// Soit α_0 le point critique de $P(\alpha)$
$\alpha_0 = 1/2(\alpha_2 - h_1/h_3)$; $f_0 = f(x^k + \alpha_0 d^k)$;
SI $f_0 < f_3$
ALORS $\alpha = \alpha_0$; ELSE $\alpha = \alpha_3$;
Fin SI;

CHAPITRE 6

ALGORITHME DU GRADIENT CONJUGUE

Méthode à pas conjugués

On s'intéresse à la recherche du minimum, $x^\star$, de la forme quadratique suivante:

$$f(x) = x^T Ax - b^T x \quad ; A = A^T > 0$$

Définition 23 (Directions conjuguées)
Deux directions d_1 et d_2 sont dites $A-$conjuguées si et seulement si :
$d_1^T A d_2 = 0$.

Proposition 11
Si un ensemble de vecteurs non nuls $\{d_1, d_2, \ldots, d_n\}$ sont mutuellement $A-$conjugués, alors ils sont linéairement indépendants.

Preuve: Supposons que ces vecteurs soient linéairement dépendants, alors il existerait un ensemble de réels $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ non tous nuls tels que:

$$\alpha_1 d_1 + \alpha_2 d_2 + \ldots + \alpha_n d_n = 0$$

Supposons par exemple que $\alpha_k \neq 0$ avec $1 \leq k \leq n$. D'après l'égalité précédente, on a:

$$d_k^T A(\alpha_1 d_1 + \alpha_2 d_2 + \ldots + \alpha_n d_n) = 0$$

mais comme les d_i sont mutuellement $A-$conjugués, on obtient:

$$\begin{aligned} d_k^T A(\alpha_1 d_1 + \alpha_2 d_2 + \ldots + \alpha_n d_n) &= \alpha_k d_k^T A d_k \\ &= 0 \end{aligned}$$

Comme $A > 0$, ceci entraine $\alpha_k = 0$, ce qui mène à une contradiction.

Pourquoi le concept de A-orthogonalité?

Supposons n vecteurs mutuellement $A-$conjugués où $A = A^T > 0$. Ces vecteurs sont linéairement indépendants et forment donc une base de $\mathbb{R}^n$. La solution $x^\star$ peut s'exprimer dans cette base comme suit:

$$x^\star = \sum_{k=1}^{n} \alpha_k d_k \qquad (\star)$$

La recherche de la solution $x^\star$ se ramène donc à celle des α_k, $k = 1, \ldots, n$.

Soit $1 \leq i \leq n$; à partir de l'équation $(\star)$ et compte tenu du fait que les vecteurs d_k sont mutuellement $A-$conjugués, nous avons:

$$d_i^T A x^\star = \alpha_i d_i^T A d_i \Longrightarrow \alpha_i = \frac{d_i^T A x^\star}{d_i^T A d_i} = \frac{d_i^T b}{d_i^T A d_i}$$

La solution $x^\star$ peut s'exprimer donc comme suit:

$$x^\star = \sum_{i=1}^{n} \frac{d_i^T b}{d_i^T A d_i} d_i$$

On remarque que les expressions des α_i sont simples et ne sont fonctions que de quantités connues, à savoir les d_i et b. De plus, la solution $x^\star$ s'améliore au fur et à mesure que l'on ajoute une composante.

Tout le problème maintenant réside dans la construction des d_i. Ce problème sera résolu ultérieurement lors de la présentation effective de l'algorithme du gradient conjugué. Nous allons maintenant supposer que l'on dispose d'un ensemble de n vecteurs $A-$ conjugués et nous allons montrer que l'on peut construire une suite de vecteurs x^k qui converge vers la solution $x^\star$.

Théorème 10 (des directions conjugués) *Soit* $\{d_0, d_2, \ldots, d_{n-1}\}$ *un ensemble de* n *vecteurs non nuls mutuellement* A*-conjugués. Alors, pour tout* $x^0 \in \mathbb{R}^n$*, la suite générée par*

$$x^{k+1} = x^k + \alpha_k d_k \qquad (\star\star)$$

avec

$$\alpha_k = -\frac{g_k^T d_k}{d_k^T A d_k} \quad et \quad g_k = Ax^k - b$$

converge vers la solution unique $x^\star$ *de* $Ax = b$ *après* n *itérations, c'est-à-dire* $x^n = x^\star$.

Preuve du théorème:

Comme les d_k forment une base de $\mathbb{R}^n$, l'élément $x^\star - x^0$ peut s'écrire comme suit:

$$\begin{aligned} x^\star - x^0 &= \alpha_0 d_0 + \ldots + \alpha_{n-1} d_{n-1} \Longrightarrow \\ \alpha_k &= \frac{d_k^T A(x^\star - x^0)}{d_k^T A d_k} \end{aligned}$$

Or, à partir du processus itératif $(\star\star)$, nous avons:

$$\begin{aligned} x^k &= x^{k-1} + \alpha_{k-1} d_{k-1} \\ &= x^{k-2} \alpha_{k-2} d_{k-2} + \alpha_{k-1} d_{k-1} \\ &\vdots \\ &= x^0 + \alpha_0 d_0 + \alpha_1 d_1 = + \ldots + \alpha_{k-1} d_{k-1} \Longrightarrow \\ x^k - x^0 &= \alpha_0 d_0 + \alpha_1 d_1 + \ldots + \alpha_{k-1} d_{k-1} \Longrightarrow \\ d_k^T A(x^k - x^0) &= 0 \quad \text{car les } d_k \text{ mutuellement } A - \text{conjugués} \end{aligned}$$

On a donc $\alpha_k = \dfrac{d_k^T A(x^\star - x^0)}{d_k^T A d_k}$ et $d_k^T A(x^k - x^0) = 0$. Il en résulte:

$$\begin{aligned}
\alpha_k &= \frac{d_k^T A(x^\star - x^0)}{d_k^T A d_k} \\
&= \frac{d_k^T A(x^\star - x^k)}{d_k^T A d_k} + \frac{d_k^T A(x^k - x^0)}{d_k^T A d_k} \\
&= \frac{d_k^T A(x^\star - x^k)}{d_k^T A d_k} \\
&= \frac{\left(A(x^\star - x^k)\right)^T d_k}{d_k^T A d_k} \\
&= \frac{\left(b - Ax^k\right)^T d_k}{d_k^T A d_k} \\
&= -\frac{g_k^T d_k}{d_k^T A d_k}
\end{aligned}$$

Méthode du gradient conjugué

Nous allons présenter dans ce qui une procédure permettant de construire n vecteurs mutuellement $A-$conjugués.

On considère la suite $\{x^k\}$ définie comme suit:

$$x^{k+1} = x^k + \alpha^k d_k$$

où x^0 est quelconque et $d_0 = -\nabla f(x^0) = -(Ax^0 - b)$.

A chaque étape, il s'agit de choisir un pas α_k et de construire un vecteur d_{k+1} qui soit conjugué à tous les d_i pour $0 \leq i \leq k$.

Dans l'algorithme du gradient conjugué, la suite des vecteurs d_k et les pas α_k sont choisis comme suit :

- Choix des d_k:

$$d_{k+1} = -\nabla f(x^{k+1}) + \beta_k d_k \quad \text{avec} \quad \beta_k = \frac{d_k^T A \nabla f(x^{k+1})}{d_k^T A d_k}$$

- Choix des α_k:

$$\alpha_k = -\frac{d_k^T \nabla f(x^k)}{d_k^T A d_k}$$

Le α_k est choisi de manière optimale pour minimiser la fonction

$$h(\alpha) = f(x^{k+1}) = f(x^k + \alpha d_k)$$

le long de la direction d_k. On a donc $h'(\alpha_k) = \nabla f(x^{k+1})^T d_k = 0$. D'où

$$\begin{aligned}
\left(Ax^{k+1} - b\right)^T d_k &= 0 \Longrightarrow \\
\left(A(x^k + \alpha_k d_k) - b)\right)^T d_k &= 0 \Longrightarrow \\
\left(Ax^k - b + \alpha_k A d_k)\right)^T d_k &= 0 \Longrightarrow \\
\left(\nabla f(x^k) + \alpha_k A d_k)\right)^T d_k &= 0 \Longrightarrow \\
(\alpha_k A d_k))^T d_k &= -\nabla f(x^k)^T d_k \Longrightarrow \\
\\
\alpha_k &= -\frac{d_k^T \nabla f(x^k)}{d_k^T A d_k}
\end{aligned}$$

Le β_k est choisi de sorte que d_{k+1} soit $A-$conjugué avec d_k.

En effet, comme $d_{k+1} = -\nabla f(x^{k+1}) + \beta_k d_k$, on a:

$$\begin{aligned} d_k^T A d_{k+1} &= -d_k^T A \nabla f(x^{k+1}) + \beta_k d_k^T A d_k \\ &= 0 \Longrightarrow \end{aligned}$$

$$\beta_k = \frac{d_k^T A \nabla f(x^{k+1})}{d_k^T A d_k}$$

Nous allons maintenant montrer que les directions d_k ainsi générées sont mutuellement $A-$conjuguées. Pour ce faire, nous allons raisonner par récurrence.

• $k = 1$:

Comme α_0 est choisi optimal pour minimiser
$f(x^1) = f(x^0 + \alpha_0 d_0) = f(x^0 - \alpha_0 \nabla f(x^0))$, on a:
$\nabla f(x^1)^T \nabla f(x^0) = 0$. Par ailleurs, β_0 a été choisi de sorte que d_1 soit $A-$ conjugué à d_0.

• Hypothèse de récurrence: $\nabla f(x^i)^T \nabla f(x^j) = d_i^T A d_j = 0$ pour tout $i, j \leq n$ avec $i \neq j$.

Il faut démontrer que:

$$\text{pour tout } \; i \leq n \; \text{ on a } \; \nabla f(x^{n+1})^T \nabla f(x^i) = d_{n+1}^T A d_i = 0$$

Notons tout d'abord l'identité suivante:

(2)$\nabla f(x^{k+1}) = \nabla f(x^k) + \alpha_k A d_k$.
En effet, on a:

$$\begin{aligned} \nabla f(x^{k+1}) &= Ax^{k+1} - b \\ &= A(x^k + \alpha_k d_k) - b \\ &= Ax^k - b + \alpha_k A d_k \\ &= \nabla f(x^k) + \alpha_k A d_k \end{aligned}$$

Montrons maintenant la récurrence , c-à-d

$$\text{pour tout } \; i \leq n \; \text{ on a } \; \nabla f(x^{n+1})^T \nabla f(x^i) = d_{n+1}^T A d_i = 0$$

Pour ce faire, on traitera tout d'abord le cas $i = n$, ensuite $i < n$.

(i) cas $i = n$:

On a: $d_{n+1} = -\nabla f(x^{n+1}) + \beta_n d_n$. On sait que le choix de β_n est fait de sorte que d_{n+1} soit $A-$ conjugué à d_n. Il nous reste maintenant à démontrer que $\nabla f(x^{n+1})^T \nabla f(x^n) = 0$. Nous avons:

$$\begin{aligned}
\left(\nabla f(x^{n+1})\right)^T \nabla f(x^n) &= \left(\nabla f(x^{n+1})\right)^T (-d_n + \beta_{n-1} d_{n-1}) \\
&= -\nabla f(x^{n+1})^T d_n + \beta_{n-1} \nabla f(x^{n+1})^T d_{n-1} \\
&= \beta_{n-1} \nabla f(x^{n+1})^T d_{n-1} \text{ car } \alpha_n \text{ est choisi de manière optimale} \\
&= \beta_{n-1} (\nabla f(x^n) + \alpha_n A d_n)^T d_{n-1} \\
&= \beta_{n-1} (\nabla f(x^n))^T d_{n-1} \\
&= 0 \text{ car } \alpha_{n-1} \text{ est choisi de manière optimale}
\end{aligned}$$

(ii) cas $i < n$:

On a d'une part:

$$\begin{aligned}
\nabla f(x^{n+1})^T \nabla f(x^i) &= \left(\nabla f(x^n) + \alpha_n A d_n)\right)^T \nabla f(x^i) \\
&= \alpha_n \left(A d_n\right)^T \nabla f(x^i) \\
&= \alpha_n \left(A d_n\right)^T \left(-d_i + \beta_{i-1} d_{i-1}\right) \\
&= 0
\end{aligned}$$

D'autre part, on a:

$$\begin{aligned}
d_{n+1}^T A d_i &= \left(-\nabla f(x^{n+1}) + \beta_n d_n\right)^T A d_i \\
&= -\left(\nabla f(x^{n+1})\right)^T A d_i \\
&= -\left(\nabla f(x^{n+1})\right)^T \left(\frac{\nabla f(x^{i+1}) - \nabla f(x^i)}{\alpha_i}\right) \\
&= 0
\end{aligned}$$

Ceci termine la démonstration.

Dans l'algorithme du gradient conjugué, nous avons: $\beta_k = \dfrac{d_k^T A \nabla f(x^{k+1})}{d_k^T A d_k}$.

Exprimons β_k autrement. D'une part, on a:

$$\begin{aligned} d_k^T A \nabla f(x^{k+1}) &= \left(\nabla f(x^{k+1})\right)^T A d_k \\ &= \left(\nabla f(x^{k+1})\right)^T \left(\frac{\nabla f(x^{k+1}) - \nabla f(x^k)}{\alpha_k}\right) \\ &= \frac{1}{\alpha_k} \left(\nabla f(x^{k+1})\right)^T \nabla f(x^{k+1}) \end{aligned}$$

D'autre part, on a:

$$\begin{aligned} d_k^T A d_k &= \left(-\nabla f(x^k) + \beta_{k-1} d_{k-1}\right)^T A d_k \\ &= -\left(\nabla f(x^k)\right)^T A d_k \\ &= -\left(\nabla f(x^k)\right)^T \left(\frac{\nabla f(x^{k+1}) - \nabla f(x^k)}{\alpha_k}\right) \\ &= \frac{1}{\alpha_k} \left(\nabla f(x^k)\right)^T \nabla f(x^k) \end{aligned}$$

Variantes de l'algorithme du gradient conjugué

$$\beta_k = \frac{d_k^T A \nabla f(x^{k+1})}{d_k^T A d_k} \text{ avec}$$

$$d_k^T A \nabla f(x^{k+1}) = \frac{1}{\alpha_k} \left(\nabla f(x^{k+1})\right)^T \nabla f(x^{k+1});\ d_k^T A d_k = \frac{1}{\alpha_k} \left(\nabla f(x^k)\right)^T \nabla f(x^k)$$

- **Variante de Fletcher & Reeves (1964)**:

$$\beta_k = \frac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2} \quad (\star)$$

Comme $\left(\nabla f(x^{k+1})\right)^T \nabla f(x^k) = 0$, l'expression $(\star)$ peut s'écrire comme suit:

- **Variante de Polak & Ribière (1969)**:

$$\beta_k = \frac{\left(\nabla f(x^{k+1}) - \nabla f(x^k)\right)^T \nabla f(x^{k+1})}{\|\nabla f(x^k)\|^2}$$

• Les expressions de β_k données par la formule originale du gradient conjugué ainsi que par les variantes de Fletcher & Reeves et Polak & Ribière sont toutes identiques lorsque le critère est quadratique. Ceci n'est plus vrai dans le cas général lorsque le critère n'est pas quadratique. De plus, dans le cas non quadratique, la convergence nécessite généralement beaucoup plus que n itérations, d'où la nécessité de réinitialiser le processus de génération des directions conjugués après n itérations en prenant $d_{n+1} = -\nabla f(x^{n+1})$ et les $(n-1)$ directions suivantes seront générés selon l'algorithme du gradient conjugué ou ses variantes.

• Lorsque le critère est non quadratique, c'est la variante de Polak & Ribière qui est la plus utilisée. Dans cette formule, on remarque que si le gradient évalué en deux point de recherche successifs ne varie beaucoup, c'est-à-dire $\nabla f(x^{k+1}) \simeq \nabla f(x^k)$, on aura $\beta_k \simeq 0$ et par conséquence $d_{k+1} \simeq -\nabla f(x^{k+1})$, ce qui peut être interprété comme une réinitialisation automatique de l'algorithme.

CHAPITRE 7

ALGORITHMES DE NEWTON ET QUASI-NEWTON

Algorithme de Newton

Pour la recherche du minimum de la fonction, l'algorithme de Newton commence par chercher une direction de descente, d_k, en considérant une approximation quadratique du critère à minimiser autour de l'estimation courante x^k. On a:

$$f(x^k + d_k) \simeq= q_k(d_k) = f(x^k) + \left(\nabla f(x^k)\right)^T d_k + \frac{1}{2} d_k^T \nabla^2 f(x^k) d_k$$

La direction de descente est calculée de façon à satisfaire le critère du premier ordre, c'est-à-dire $\nabla q_k(d_k) = 0$.

Or, comme $\nabla q(d_k) = \nabla^2 f(x^k) d_k + \nabla f(x^k)$, la direction de descente d_k est solution du système:

$$\nabla^2 f(x^k) d_k + \nabla f(x^k) = 0$$

Après la recherche de la direction de descente, la nouvelle estimée x^{k+1} est calculée comme suit:

$$x^{k+1} = x^k + \alpha_k d_k$$

où α_k est le pas descente qui peut être calculé comme dans les algorithmes du gradient.

On remarque donc que la direction d_k calculée par l'algorithme de Newton n'est une direction de descente que si la matrice hessienne, $\nabla^2 f(x^k)$ est définie positive.

On remarque aussi que l'algorithme de Newton nécessite l'évaluation de la hessienne à chaque itération. Un tel calcul est coûteux et il serait donc intéressant de réfléchir à approcher cette matrice (ou encore mieux son inverse) à partir des informations fournies par $\nabla f(x^k)$. Ceci nous conduit aux méthodes de Quasi-Newton.

Méthodes de Quasi-Newton

• **Généralités**

Il s'agit d'imiter l'algorithme de Newton sans calculer la matrice hessienne, ni son inverse. L'algorithme itératif de Quasi-Newton s'écrit comme suit:

$$x^{k+1} = x^k - \alpha_k S_k \nabla f(x^k)$$

où S_k est une matrice symétrique définie positive qui est une approximation de $\left(\nabla^2 f(x^k)\right)^{-1}$ et α_k est un réel positif fourni par une procédure de recherche linéaire le long de la direction $d_k = -S_k \nabla f(x^k)$.

Il est clair que, plus S_k est proche (au sens d'une norme matricielle) de $\left(\nabla^2 f(x^k)\right)^{-1}$, plus l'algorithme convergera rapidement.

• **Algorithmes de Quasi-Newton:**

$$x^{k+1} = x^k - \alpha_k S_k \nabla f(x^k)$$

A l'étape k, la mise à jour de la matrice S_k se fait à l'aide d'une formule additive simple :

$$S_{k+1} = S_k + C_k$$

où C_k est matrice de correction qui intègre au mieux la nouvelle information fournie par x^{k+1} et $\nabla f(x^{k+1})$. En particulier, C_k sera choisie de telle manière que S_{k+1} satisfasse une équation dite "équation de la sécante" ou encore "condition de Quasi-Newton":

$$S_{k+1}\left(\nabla f(x^{k+1}) - \nabla f(x^k)\right) = x^{k+1} - x^k$$

Pourquoi la condition de Quasi-Newton?

Supposons que $f \in \mathcal{C}^2$ et effectuons un développement de Taylor de ∇f au voisinage de x^k:

$$\begin{aligned} \nabla f(x) \simeq \nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) \quad &\Longrightarrow \\ \left(\nabla^2 f(x^k)\right)^{-1} \left(\nabla f(x) - \nabla f(x^k)\right) \quad &\simeq \quad x - x^k \end{aligned}$$

Cette approximation est exacte si le critère f est quadratique. Pour $x = x^{k+1}$, si S_k est une bonne approximation de $\left(\nabla^2 f(x^k)\right)^{-1}$, cette équation devient:

$$S_k \left(\nabla f(x^{k+1}) - \nabla f(x^k)\right) \quad \simeq \quad x^{k+1} - x^k \quad (\star)$$

Mais, comme x^{k+1} est calculé après S_k, l'équation $(\star)$ ne peut pas être satisfaite même approximativement. Par contre, on peut toujours imposer à S_{k+1} de satisfaire exactement $(\star)$ où S_k est remplacé par S_{k+1}. On obtient alors la condition de Quasi-Newton.

L'objectif consiste donc à trouver une suite de matrices S_k, simple à construire, n'utilisant que l'information fournie par $\nabla f(x^k)$, qui converge vite vers des approximations de plus en plus pécises de l'inverse de la matrice hessienne de f, et qui satisfait la condition de Quasi-Newton.

Exemple: Soit f un critère quadratique de hessien $A > 0$ d'ordre n. Considérons une famille de vecteurs $(d_1, d_2, \ldots, d_n)$ de vecteurs conjugués deux à deux par rapport à A et définissons la matrice S_k comme suit:

$$S_k = \sum_{i=1}^{k} \frac{d_i di^T}{d_i^T A d_i} = \sum_{i=1}^{k} \frac{d_i di^T}{\|d_i\|_A}$$

nous remarquons que la suite S_k est construire à partir d'une formule de type $S_{k+1} = S_k + C_k$.

Montrons que $S_n^{-1} = A$. En effet, soit $k \in \{1, \ldots, n\}$; on a:

$$S_n A d_k = \sum_{i=1}^{k} \frac{d_i di^T A d_k}{\|d_i\|_A} = d_k$$

Pour tout $k \in \{1, \ldots, n\}$, d_k est un vecteur propre de $S_n A$ qui est associé à la valeur propre $S_n A$. Comme cette matrice est d'ordre n et qu'elle a n valeurs propres toutes égales à 1, elle est égale à la matrice identité et on a donc $S_n = A^{-1}$.

On voit à travers cet exemple que l'on peut construire en n itérations l'inverse d'une matrice symétrique définie positive. Dans la suite, nous allons adopter les notations suivantes:

$$\gamma_k = \nabla f(x^{k+1}) - \nabla f(x^k) \quad ; \quad \delta_k = x^{k+1} - x^k$$

Avec ces notations, la condition de Quasi-Newton devient:

$$S_{k+1} \gamma_k = \delta_k$$

Correction de rang 1

Comme $\left(\nabla^2 f(x^k)\right)^{-1}$ est symétrique, le terme de correction C_k doit être choisi symétrique. On choisit la mise à jour suivante de S_k:

$$S_{k+1} = S_k + a_k v_k v_k^T$$

où $v_k \in \mathbb{R}^n$ et $a_k \in \mathbb{R}$ sont choisis pour satisfaire la condition de Quasi-Newton, en l'occurrence:

$$\begin{aligned} S_{k+1}\gamma_k = \delta_k \quad &\Longrightarrow \quad S_k\gamma_k + a_k v_k v_k^T \gamma_k = \delta_k \Longrightarrow \\ a_k v_k v_k^T \gamma_k = \delta_k - S_k\gamma_k \quad &\Longrightarrow \quad a_k v_k^T \gamma_k v_k = \delta_k - S_k\gamma_k \end{aligned}$$

Il suffit donc de choisir v_k et a_k comme suit :

$$v_k = \delta_k - S_k\gamma_k \quad \text{et} \quad a_k = \frac{1}{v_k^T\gamma_k} = \frac{1}{\left(\delta_k - S_k\gamma_k\right)^T \gamma_k}$$

Dans la plupart des méthodes de Quasi-Newton, lorsque la condition de Quasi-Newton est vérifiée, on a aussi:

$$S_{k+1}\gamma_i = \delta_i \quad \text{pour} \quad i \leq k \quad (\star)$$

De ce fait, supposons que l'on ait n vecteurs δ_k, $k = 0, \ldots, n-1$, indépendants et posons:

$$\Gamma = [\gamma_0 \ \gamma_1 \ldots \ \gamma_{n-1}]$$
$$\Delta = [\delta_0 \ \delta_1 \ldots \ \delta_{n-1}]$$

Si l'égalité $(\star)$ est vérifiée, nous aurons $S_n\Gamma = \Delta$. Comme Δ est inversible (puisque les vecteurs δ_k sont indépendants), S_n et Γ le sont aussi et on a: $\Gamma^{-1}S_n^{-1} = \Delta^{-1} \Longrightarrow S_n^{-1} = \Gamma\Delta^{-1}$.

Nous allons montrer dans ce qui suit que si le critère f est quadratique de matrice $A = A^T > 0$, alors on aura $A = \Gamma\Delta^{-1} = S_n^{-1}$. C'est que nous énonçons au théorème suivant.

Théorème 11 méthode de correction de rang 1:
Si f est quadratique de hessien A et si δ_0, δ_1, ..., δ_{n-1} sont n vecteurs linéairement indépendants, alors la méthode de correction de rang 1 converge en n itérations et on a $S_n^{-1} = A$.

Preuve: il suffit de montrer que les matrices S_k, $k = 1, \ldots, n-1$, construites par

$$S_{k+1} = S_k + a_k v_k v_k^T$$

satisfont la condition

$$S_{k+1}\gamma_i = \delta_i \quad \text{pour} \quad i \leq k$$

Pour ce faire, nous allons raisonner par récurrence.

• Pour $k = 0$, nous devons montrer que $S_1\gamma_0 = \delta_0$. Or cette égalité est la condition de Quasi-Newton pour S_1 et elle est donc vérifiée.

• Pour l'ordre, k, nous avons d'après l'hypothèse de récurrence

$$S_k\gamma_i = \delta_i \quad \text{pour} \quad i \in \{0, \ldots, k-1\}$$

et nous devons donc montrer

$$S_{k+1}\gamma_i = \delta_i \quad \text{pour} \quad i \in \{0, \ldots, k\}$$

Soit $i \in \{0, \ldots, k\}$. Nous allons tout d'abord traiter le cas $i = k$, ensuite le cas $i \in \{0, \ldots, k-1\}$.

- $i = k$: il s'agit de montrer que $S_{k+1}\gamma_k = \delta_k$. Or cette égalité est la condition de Quasi-Newton pour S_{k+1} et elle est donc vérifiée.

- $i \in \{0, \ldots, k-1\}$: nous avons

$$S_{k+1}\gamma_i = S_k\gamma_i + \frac{(\delta_k - S_k\gamma_i)\,(\delta_k - S_k\gamma_k)^T}{(\delta_k - S_k\gamma_k)^T\,\gamma_k}\gamma_i$$

Or, nous avons $(\delta_k - S_k\gamma_k)^T\,\gamma_i$ $= \delta_k^T\gamma_i - \gamma_k^T S_k\gamma_i$ (car S_k est symétrique)

$= \delta_k^T\gamma_i - \gamma_k^T\delta_i$ (par hypothèse de récurrence)

$= \delta_k^T\gamma_i - \delta_i^T\gamma_k$

Par ailleurs, par définition de A, nous avons:

$$A(x^{k+1} - x^k) = \nabla f(x^{k+1}) - \nabla f(x^k) \;\text{ c-à-d }\; A\delta_k = \gamma_k$$

Il en résulte

$$(\delta_k - S_k\gamma_k)^T\,\gamma_i = \delta_k^T\gamma_i - \delta_i^T A\delta_k = \delta_k^T\gamma_i - \delta_k^T A\delta_i = \delta_k^T\gamma_i - \delta_k^T\gamma_i = 0$$

Ceci termine la démonstration.

Algorithme de Quasi-Newton avec correction de rang 1

• on se donne x^0 et on pose $S_0 = I_n$.

• à l'itération k: $$x^{k+1} = x^k - \alpha_k S_k \nabla f(x^k)$$

où α_k est obtenu par une recherche linéaire le long de $(-S_k \nabla f(x^k)$. La matrice S_k est mise à jour comme suit :

$$S_{k+1} = S_k + \frac{(\delta_k - S_k\gamma_i)\,(\delta_k - S_k\gamma_k)^T}{(\delta_k - S_k\gamma_k)^T\,\gamma_k}$$

où $\delta_k = x^{k+1} - x^k$ et $\gamma_k = \nabla f(x^{k+1}) - \nabla f(x^k)$.

Remarque: le caractère défini positif de la matrice S_k n'est pas assuré par la méthode. Ceci est dû au fait que le coefficient $a_k = (\delta_k - S_k\gamma_k)^T\,\gamma_k$ peut devenir très petit, voir négatif et la matrice S_k peut perdre son caractère de définie positive même si S_{k-1} l'était.

Correction de rang 2

Nous allons présenter deux méthodes utilisant une correction de rang 2:

- la méthode de Davidson-Fletcher-Powell (D.F.P)

- la méthode de Broyden-Goldfarb-Shanno (BFGS).

Algorithme de D.F.P.

Appliquée à un critère quadratique de hessienne A, cette méthode calcule une approximation de l'inverse de la hessienne tout en engendrant des directions δ_i qui sont mutuellement A-conjuguées.

La matrice S_k est mise à jour comme suit:

$$S_{k+1} = S_k + a_k v_k v_k^T + b_k w_k w_k^T$$

où $v_k, w_k \in \mathbb{R}^n$ et $a_k, b_k \in \mathbb{R}$ sont choisis pour satisfaire la condition de Quasi-Newton, en l'occurrence:

$$\begin{aligned}
S_{k+1}\gamma_k = \delta_k \quad &\Longrightarrow \\
S_k\gamma_k + a_k v_k v_k^T \gamma_k + b_k w_k w_k^T \gamma_k = \delta_k \quad &\Longrightarrow \\
a_k v_k v_k^T \gamma_k + b_k w_k w_k^T \gamma_k = \delta_k - S_k\gamma_k \quad &\Longrightarrow \\
a_k v_k^T \gamma_k v_k + b_k w_k^T \gamma_k w_k = \delta_k - S_k\gamma_k &
\end{aligned}$$

• Condition de Quasi-Newton : $a_k v_k^T \gamma_k v_k + b_k w_k^T \gamma_k w_k = \delta_k - S_k \gamma_k$

Pour satisfaire cette condition, il suffit de prendre:

$$\begin{aligned} a_k v_k^T \gamma_k &= 1 \\ v_k &= \delta_k \\ b_k w_k^T \gamma_k &= -1 \\ w_k &= S_k \gamma_k \end{aligned}$$

La mise à jour de $S_{k+1} = S_k + a_k v_k v_k^T + b_k w_k^T w_k$ devient alors:

$$S_{k+1} = S_k + \frac{\delta_k \delta_k^T}{\delta_k^T \gamma_k} - \frac{S_k \gamma_k \gamma_k^T S_k}{\gamma_k^T S_k \gamma_k}$$

Nous allons maintenant résumer les propriétés de cette méthode dans le théorème qui suit.

Théorème 12 Algorithme de D.F.P.:
L'algorithme de D.F.P. a les trois propriétés suivantes:

1) A l'étape k, si $S_k = S_k^T > 0$ et si la recherche linéaire est exacte (ou encore si $\delta_k^T \gamma_k > 0$), alors $S_{k+1} = S_{k+1}^T > 0$.

2) Si le critère f est quadratique de hessienne $A = A^T > 0$, alors l'algorithme de D.F.P. est tel que:

pour tout $0 \leq i < j \leq k : \delta_i^T A \delta_j = 0$ *et pour tout* $0 \leq i \leq k : S_{k+1} A \delta_i = \delta_i$

3) Les deux propriétés précédentes impliquent dans le cas quadratique la convergence de la méthode en n itérations, c'est-à-dire $S_n = A^{-1}$.

Preuve: soit $u \in \mathbb{R}^n$ un vecteur non nul. Nous avons

$$\begin{aligned} u^T S_{k+1} u &= u^T S_k u + \frac{u^T \delta_k \delta_k^T u}{\delta_k^T \gamma_k} - \frac{u^T S_k \gamma_k \gamma_k^T S_k u}{\gamma_k^T S_k \gamma_k} \\ &= u^T S_k u + \frac{\left(u^T \delta_k\right)^2}{\delta_k^T \gamma_k} - \frac{\left(u^T S_k \gamma_k\right)^2}{\gamma_k^T S_k \gamma_k} \\ &= \frac{(u^T S_k u)(\gamma_k^T S_k \gamma_k) - \left(u^T S_k \gamma_k\right)^2}{\gamma_k^T S_k \gamma_k} + \frac{\left(u^T \delta_k\right)^2}{\delta_k^T \gamma_k} \end{aligned}$$

D'une part, nous avons d'après l'inégalité de Cauchy-Schwartz:

$$(u^T S_k u)(\gamma_k^T S_k \gamma_k) = \|u\|_{S_k}^2 \|\gamma_k\|_{S_k}^2 \geq \left(u^T S_k \gamma_k\right)^2 \quad (\star)$$

et l'inégalité n'a lieu que si u et γ_k sont colinéaires.
D'autre part, nous avons $x^{k+1} = x^k - \alpha_k S_k \nabla f(x^k)$. La recherche linéaire étant exacte, nous obtenons:

$$\left(\nabla f(x^{k+1})\right)^T S_k \nabla f(x^k) = 0$$

Il s'ensuit:

$$\begin{aligned}
\delta_k^T \gamma_k &= \left(x^{k+1} - x^k\right)^T \left(\nabla f(x^{k+1}) - \nabla f(x^k)\right) \\
&= \left(-\alpha_k S_k \nabla f(x^k)\right)^T \left(\nabla f(x^{k+1}) - \nabla f(x^k)\right) \\
&= \alpha_k \left(\nabla f(x^k)\right)^T S_k \nabla f(x^k) \\
&> 0 \quad \text{tant que } x^k \text{ n'est pas l'optimum}
\end{aligned}$$

Pour montrer que le membre gauche de l'égalité $(\star)$ est strictement positif, il reste à démontrer que les deux termes du membre droit de cette égalité ne peuvent pas s'annuler simultanément. En effet, le numérateur du premier membre est nul si et seulement si u et γ_k sont colinéaires c-à-d $u = a\gamma_k$ où a est un réel strictement positif (car u est non nul). Dans ce cas, le deuxième terme s'écrit:

$$\frac{\left(u^T \delta_k\right)^2}{\gamma_k^T \delta_k} = a^2 \gamma_k^T \delta_k > 0 \quad \text{puisque } a > 0$$

2) Nous allons procéder à un raisonnement par récurrence. Notons tout d'abord l'identité suivante:

$$\gamma_k = \nabla f(x^{k+1}) - \nabla f(x^k) = A(x^{k+1} - x^k) = A\delta_k$$

• Pour commencer la récurrence, il faut montrer que $S_1 A\delta_0 = \delta_0$ et $\delta_0^T A\delta_1 = 0$
D'une part, nous avons:

$$S_1 A\delta_0 = S_1\gamma_0 = \delta_0$$

D'autre part, nous avons:

$$\begin{aligned}
\delta_0^T A\delta_1 &= \delta_0^T A(-\alpha_1 S_1 \nabla f(x^1)) \\
&= -\alpha_1 \left(S_1 A\delta_0\right)^T \nabla f(x^1) \\
&= -\alpha_1 \delta_0^T \nabla f(x^1) \\
&= -\alpha_1 (-\alpha_0 S_0 \nabla f(x^0))^T \nabla f(x^1) \\
&= 0 \quad \text{car } \alpha_0 \text{ est choisi de manière optimale}
\end{aligned}$$

- Nous supposons maintenant que l'on a:

$$\text{pour tout } 0 \leq i < j \leq k-1 : \ \delta_i^T A \delta_j = 0$$

$$\text{pour tout } 0 \leq i \leq k-1 : S_k A \delta_i = \delta_i$$

Nous devons démontrer:

$$\text{pour tout } 0 \leq i < j \leq k : \ \delta_i^T A \delta_j = 0$$

$$\text{pour tout } 0 \leq i \leq k : S_{k+1} A \delta_i = \delta_i$$

Nous allons commencer par démontrer que pour tout $0 \leq i < j \leq k$, nous avons $\delta_i^T A \delta_j = 0$. Pour ce faire il suffit de considérer le cas où $j = k$ car pour $j \leq k-1$ le résultat est vrai d'après l'hypothèse de récurrence. Nous allons tout d'abord considérer le cas $i = k-1$, puis $i < k-1$.

Soit $i = k - 1$. Nous avons:

$$\begin{aligned}
\delta_{k-1}^T A \delta_k &= \gamma_{k-1}^T \delta_k \\
&= \gamma_{k-1}^T (-\alpha_k S_k \nabla f(x^k)) \\
&= -\alpha_k \left(S_k \gamma_{k-1}\right)^T \nabla f(x^k) \\
&= -\alpha_k \delta_{k-1}^T \nabla f(x^k) \\
&= -\alpha_k \left(-\alpha_{k-1} S_{k-1} \nabla f(x^{k-1})\right)^T \nabla f(x^k) \\
&= 0 \text{ car } \alpha_{k-1} \text{ est choisi de manière optimale}
\end{aligned}$$

Soit maintenant $i < k-1$. Nous avons:

$$\begin{aligned}
\delta_i^T A\delta_k &= \delta_i^T A(-\alpha_k S_k \nabla f(x^k)) \\
&= -\alpha_k \left(S_k A\delta_i\right)^T \nabla f(x^k) \\
&= -\alpha_k \delta_i^T \nabla f(x^k) \\
&= -\alpha_k \delta_i^T (\gamma_{k-1} + \nabla f(x^{k-1})) \\
&= -\alpha_k \delta_i^T (A\delta_{k-1} + \nabla f(x^{k-1})) \\
&= -\alpha_k \delta_i^T \nabla f(x^{k-1}) \\
&= -\alpha_k \delta_i^T \left(\frac{-1}{\alpha_{k-1}} S_{k-1}^{-1}\delta_{k-1}\right) \\
&= \frac{\alpha_k}{\alpha_{k-1}} \delta_i^T S_{k-1}^{-1}\delta_{k-1} \\
&= \frac{\alpha_k}{\alpha_{k-1}} \delta_i^T A\delta_{k-1} \\
&= 0
\end{aligned}$$

Pour finir, il reste à démontrer que pour tout $0 \leq i \leq k$, nous avons $S_{k+1} A\delta_i = \delta_i$. Nous allons tout d'abord considérer le cas $i = k-1$, ensuite $i < k-1$. Nous avons: $S_{k+1} A\delta_k = S_{k+1}\gamma_k = \delta_k$ d'après l'égalité de Quasi-Newton.

Soit maintenant $i < k-1$. Nous avons:

$$\begin{aligned}
S_{k+1} A\delta_i &= \left(S_k + \frac{\delta_k \delta_k^T}{\delta_k^T \gamma_k} - \frac{S_k \gamma_k \gamma_k^T S_k}{\gamma_k^T S_k \gamma_k} \right) A\delta_i \\
&= S_k A\delta_i - \frac{\delta_k \delta_k^T A\delta_i}{\delta_k^T \gamma_k} - \frac{S_k \gamma_k \gamma_k^T S_k A\delta_i}{\gamma_k^T S_k \gamma_k} \\
&= \delta_i - \frac{\delta_k \delta_k^T A\delta_i}{\delta_k^T \gamma_k} - \frac{S_k \gamma_k \gamma_k^T \delta_i}{\gamma_k^T S_k \gamma_k} \\
&= \delta_i - \frac{\delta_k \delta_k^T A\delta_i}{\delta_k^T \gamma_k} - \frac{S_k \gamma_k \delta_k^T A\delta_i}{\gamma_k^T S_k \gamma_k} = \delta_i
\end{aligned}$$

La dernière égalité résulte du fait que $\delta_k^T A\delta_i = 0$ à partir de l'hypothèse de récurrence.

3) - Les n directions δ_i engendrées par la méthode sont conjuguées. Elles sont donc linéairement indépendantes et forment une base de $\mathbb{R}^n$. Soit Δ la matrice carrée : $\Delta = [\delta_0\ \delta_1\ \ldots\ \delta_{n-1}]$. Cette matrice est donc inversible.

- La matrice $S_n A$ admet n vecteurs propres δ_i, $i = 0, \ldots, n-1$ qui sont tous associés à la valeur propre 1. Nous avons donc $S_n A\Delta = \Delta$ et par conséquent $S_n = A^{-1}$.

Remarques:
1) L'algorithme D.F.P. est sensible à la précision de la recherche linéaire. Sa mise en oeuvre est précédée généralement d'un certain nombre de procédures de dimensionnement pour améliorer le rapport de grandeur entre les deux corrections d'ordre 1.

2) Les méthodes de Quasi-Newton, comme la méthode de Newton, ne sont pas toujours des méthodes de descente et il faudrait réinitialiser régulièrement S_k.

Algorithme de B.F.G.S.

La mise à jour de la matrice S_k se fait selon la formule suivante:

$$S_{k+1} = S_k + \left(1 + \frac{\gamma_k^T S_k \gamma_k}{\delta_k^T \gamma_k}\right) \frac{\delta_k \delta_k^T}{\delta_k^T \gamma_k} - \frac{\delta_k \gamma_k^T S_k + S_k \gamma_k \delta_k^T}{\delta_k^T \gamma_k}$$

Cette formule de mise à jour est moins sensible aux erreurs de recherche linéaire que celle de D.F.P. et elle est considérée comme étant la meilleure actuellement. On peut la retrouver à partir D.F.P. comme suit.

Pour D.F.P., nous avons:

$$S_{k+1} = S_k + \frac{\delta_k \delta_k^T}{\delta_k^T \gamma_k} - \frac{S_k \gamma_k \gamma_k^T S_k}{\gamma_k^T \gamma_k}$$

On pose $B_k = S_k^{-1}$ et on intervertit entre eux δ_k et γ_k ($S_{k+1}\gamma_k = \delta_k \Longrightarrow B_{k+1}\delta_k = \gamma_k$). On obtient alors

$$B_{k+1} = B_k + \frac{\gamma_k \gamma_k^T}{\gamma_k^T \delta_k} - \frac{B_k \delta_k \delta_k^T B_k}{\delta_k^T B_k \delta_k}$$

Il suffit d'appliquer maintenant la formule d'inversion matricielle à D.F.P. pour retrouver B.F.G.S.

Le problème des Moindres Carrés Non Linéaires (MCNL)

On dispose d'un modèle paramétrique $z = f(x, y)$ décrivant les variations d'une grandeur réelle z en fonction d'une ou de plusieurs grandeurs rassemblées au sein du vecteur y et d'un vecteur de paramètres inconnus x.

En supposant connu un échantillon (z_1, y_1), (z_2, y_2), ..., (z_M, y_M) de couples de valeurs de z et de y, l'objectif est d'estimer le vecteur des paramètres $x \in \mathbb{R}^N$ en minimisant la somme des carrés entre les valeurs observées et les valeurs prévues par le modèle:

$$S(x) = \sum_{j=1}^{M} \left(z_j - f(x, y_j)\right)^2$$

Exemple 1: la loi de Monod : $\mu = \dfrac{\mu_{max} C}{K_C + C}$ fournit un modèle pour le taux spécifique de croissance d'une biomasse consommant un substrat C.

Le vecteur des paramètres est $x = (\mu_{max}, K_C)^T$ où μ_{max} est le taux spécifique maximum de croissance et K_C est la constante de saturation.

En supposant qu'un échantillon de mesures $(\mu_1, C_1), (\mu_2, C_2), \ldots, (\mu_M, C_M)$ soit disponible, le critère s'écrit dans ce cas comme suit:

$$S(\mu_{max}, K_C) = \sum_{j=1}^{M} \left(\mu_j - \frac{\mu_{max} C_j}{K_C + C_j} \right)^2$$

Exemple 2: la loi logistique : $z = \dfrac{a}{b + ce^{-at}}$ fournit un modèle pour calculer la population mondiale en fonction du temps t.

Le vecteur des paramètres est $x = (a, b, c)^T$.

En supposant que le nombre de la population mondiale, z_j est connu à l'instant t_j pour $j = 1, \ldots, M$, le critère devient dans ce cas:

$$S(a, b, c) = \sum_{j=1}^{M} \left(z_j - \frac{a}{b + ce^{-at_j}} \right)^2$$

Analyse du problème des MCNL

Pour tout $j \in \{1, \ldots, M\}$, posons $f_j(x) = f(x, z_j)$. Le critère $S(x)$ se récrit comme suit:

$$S(x) = \sum_{j=1}^{M} \left(z_j - f_j(x)\right)^2$$

Le gradient et la hessienne du critère S sont alors donnés par les formules:

$$\begin{cases} \nabla S(x) = 2 \displaystyle\sum_{j=1}^{M} \left(f_j(x) - z_j\right) \nabla f_j(x) \\ \nabla^2 S(x) = 2 \displaystyle\sum_{j=1}^{M} \left(f_j(x) - z_j\right) \nabla^2 f_j(x) + 2 \sum_{j=1}^{M} \nabla f_j(x) \left(\nabla f_j(x)\right)^T \end{cases}$$

On note que la hessienne $\nabla^2 S(x)$ est la somme de deux termes $H_1(x)$ et $H_2(x)$:

$$\begin{cases} H_1(x) = 2\sum_{j=1}^{M} \nabla f_j(x) \left(\nabla f_j(x)\right)^T \\ H_2(x) = 2\sum_{j=1}^{M} \left(f_j(x) - z_j\right) \nabla^2 f_j(x) \end{cases}$$

La matrice $H_1(x)$ est toujours semi-définie positive et elle sera définie positive dès que $\nabla f_1(x), \nabla f_2(x), \ldots, \nabla f_M(x)$ forme une famille génératrice de $\mathbb{R}^n$, ce qui est d'autant plus probable que M est grand. En effet, soit $u \in \mathbb{R}^n$; on a:

$$u^T H_1(x) u = 2u^T \sum_{j=1}^{M} \nabla f_j(x) \left(\nabla f_j(x)\right)^T u = 2\sum_{j=1}^{M} \| \left(\nabla f_j(x)\right)^T u \| \geq 0$$

En ce qui concerne la matrice $H_2(x)$, on remarque qu'on peut la négliger devant $H_1(x)$ lorsque les résidus $r_j = f_j(x) - z_j$ sont petits et l'on peut alors confondre $\nabla^2 S(x)$ avec $H_1(x)$.

Le problème des MCNL peut se reformuler de manière matricielle comme suit.

Posons: $R : \mathbb{R}^n \to \mathbb{R}^M$, $x \mapsto R(x) = \begin{pmatrix} R_1(x) \\ R_2(x) \\ \vdots \\ R_M(x) \end{pmatrix} = \begin{pmatrix} f_1(x) - z_1 \\ f_2(x) - z_2 \\ \vdots \\ f_M(x) - z_M \end{pmatrix}$. Soit

$J(x)$ la matrice jacobienne de R au point x. On a:

$$J(x) = \begin{pmatrix} (\nabla f_1(x))^T \\ (\nabla f_2(x))^T \\ \vdots \\ (\nabla f_M(x))^T \end{pmatrix} \Longrightarrow (J(x))^T = (\nabla f_1(x)\ \nabla f_2(x)\ \ldots\ \nabla f_M(x))$$

Il s'ensuit:

$$\begin{aligned} S(x) &= (R(x))^T R(x) = \|R(x)\|^2 \\ \nabla S(x) &= 2\,(J(x))^T R(x);\quad \nabla^2 S(x) \simeq H_1(x) = 2\,(J(x))^T J(x) \end{aligned}$$

Algorithme de Gauss-Newton

Dans le cas du problème des MCNL, l'algorithme de Newton est connu aussi sous le nom d'algorithme de Gauss-Newton. Il peut s'écrire comme suit:

- $x \leftarrow x_0$, $R \leftarrow R(x_0)$, $J \leftarrow J(x_0)$

- Tant que $\|J^T R\| \geq TOL$ faire

Calculer le pas optimal t dans la direction de $u = -(J^T J)^{-1} J^T R$
$x \leftarrow x + tu$, $R \leftarrow R(x)$, $J \leftarrow J(x)$

/* fin tant que */

- Retourner x

CHAPITRE 8

OPTIMISATION SOUS CONTRAINTES D'EGALITE

Formulation du problème

On s'intéresse à des problèmes d'optimisation sous contraintes d'égalité dans $\mathbb{R}^n$.

• Ω un ouvert de $\mathbb{R}^n$, f et $g_1, \ldots, g_m$ des fonctions définies sur Ω à valeurs dans $\mathbb{R}$ et $(c_1, \ldots, c_m) \in \mathbb{R}^m$.

Le problème considéré consiste à chercher

$$\inf_{x \in A} f(x) \tag{1}$$

avec

$$A = \{x \in \Omega : \ g_j(x) = c_j, \ j = 1, \ldots, m\}$$

La fonction f est appelée fonction objectif ou fonction coût. Les fonctions g_j et

les réels c_j définissent les contraintes d'égalité. Les éléments de A s'appellent les éléments admissibles.

Quelques résultats d'algèbre

Soit E un $\mathbb{R}-$espace vectoriel et soient E_1 et E_2 deux s.e.v. de E. On dit que E_1 et E_2 sont supplémentaires, et on note $E = E_1 \oplus E_2$, si et seulement si pour tout $x \in E$, il existe un unique $(x_1, x_2) \in E_1 \times E_2$ tel que $x = x_1 + x_2$.

Proposition 12 *Soient E et F deux $\mathbb{R}-$ e.v., $v \in L(E, F)$ (ensemble des applications linéaires de E dans F), $E_1 = ker(v)$ et E_2 un supplémentaire de E_1 dans E. Alors, l'application $w = v_{|E_2} : E_2 \to Im(v)$ est un isomorphisme.*

Preuve:

Il s'agit de montrer que l'application w est à la fois surjective et injective.

"surjectivité": soit $y \in Im(v)$, il existe donc $x \in E$ tel que $y = v(x)$. Comme $E = E_1 \oplus E_2$, il existe un unique $(x_1, x_2) \in E_1 \times E_2$ tel que $x = x_1 + x_2$. Par définition de E_1, on a $v(x_1) = 0$, d'où $y = v(x_1 + x_2) = v(x_2) = w(x_2) = 0$ et w est donc surjective.

"injectivité": Soit $x \in E_2$ tel que $w(x) = 0 = v(x)$. On a donc $x \in E_1$ et par conséquent $x \in E_1 \cap E_2 = \{0\}$, d'où w est injective.

Variante du lemme de Farkas

Lemme 4 (admis) *Soit E un $\mathbb{R}-$ e.v., $u_1, \ldots, u_m$ et v, $(m+1)$ formes linéaires sur E. Les deux assertions suivantes sont équivalentes:*

1- $\bigcap_{j=1}^{m} ker\ u_j \subset ker\ v,$

2- Il existe $(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$ *telle que:*

$$v = \sum_{j=1}^{m} \lambda_j u_j$$

Conditions du premier ordre de Lagrange

Proposition 13 *Soit $x^\star \in A$ une solution locale de (1). On suppose que:*

1- f est différentiable en $x^\star$

2- g est de classe C^1 au voisinage de $x^\star$

3- $\nabla g_1(x^\star), \ldots, \nabla g_m(x^\star)$ sont linéairement indépendants

alors pour tout $h \in I\!R^n$, on a:

$$g'(x^\star)(h) = 0 \Longrightarrow (\nabla f(x^\star))^T .h = 0 \qquad (2)$$

Preuve:

Il s'agit de montrer que :

$$E_1 = ker(g'(x^\star)) = \bigcap_{i=1}^{m} ker(g'_j(x^\star)) \subset ker(f'(x^\star)) = \nabla f(x^\star)^\perp$$

Soit E_2 un supplémentaire de E_1 dans $\mathbb{R}^n$. Dans la suite, on notera tout $x \in \mathbb{R}^n$ comme $x = (x_1, x_2)$ avec $x_1 \in E_1$ et $x_2 \in E_2$. De même, on notera ∂_i, $i = 1, 2$ les différentielles partielles selon la partition $R^n = E_1 \oplus E_2$. Notons que:

$$\partial_1 g(x^\star) = g'(x^\star)_{|E_1} = 0$$

D'après la proposition (12), $\partial_2 g(x^\star)(= g'(x^\star)_{|E_2})$ est un isomorphisme de E_2 sur $Im(g'(x^\star))$. Mais, comme $\nabla g_1(x^\star), \ldots, \nabla g_m(x^\star)$ sont linéairement indépendants, $g'(x^\star)$ est surjective et nous avons donc $Im(g'(x^\star)) = \mathbb{R}^m$. Il s'ensuit que $\partial_2 g(x^\star)$ est un isomorphisme de E_2 sur $\mathbb{R}^m$.

Comme $g(x^\star) - c = 0$ et $\partial_2 g(x^\star)$ est inversible, il résulte du théorème des fonctions implicites qu'il existe un voisinage ouvert V_1 de $x_1^\star$, un voisinage ouvert V_2 de $x_2^\star$ et une application $\varphi \in C^1(V_1, V_2)$ tels que $V_1 \times V_2 \subset \Omega$, $x_2^\star = \varphi(x_1^\star)$ et

$$A \cap (V_1 \times V_2) = \{(x_1, \varphi(x_1)) : \ x_1 \in V_1\}.$$

Pour $x_1 \in V_1$, considérons l'équation suivante valable sur V_1:

$$g(x_1, \varphi(x_1)) = c \tag{3}$$

Dérivons (3) par rapport à x :

$$\partial_1 g(x_1, \varphi(x_1)) + \partial_2 g(x_1, \varphi(x_1)) \circ \varphi'(x_1) = 0 \ \forall x_1 \in V_1 \tag{4}$$

$$\partial_1 g(x_1, \varphi(x_1)) + \partial_2 g(x_1, \varphi(x_1)) \circ \varphi'(x_1) = 0 \ \forall x_1 \in V_1 \tag{5}$$

En prenant $x_1 = x_1^\star$ et compte tenu des faits que $\partial_1 g(x^\star) = 0$ et que $\partial_2 g(x_1^\star, \varphi(x_1^\star)) = \partial_2 g(x^\star)$ est inversible, on obtient:

$$\varphi'(x_1^\star) = 0$$

Soit $h \in E_1$, pour $t \in \mathbb{R}$ suffisamment petit, on a $x_1^\star + th \in V_1$, $(x_1^\star + th, \varphi(x_1^\star + th)) \in A$ et:

$$f(x_1^\star + th, \varphi(x_1^\star + th)) \geq f(x^\star) = f(x_1^\star, \varphi(x_1^\star)).$$

Considérons maintenant la fonction de la variable réelle, t, définie sur un voisinage ouvert de 0, à valeurs dans $\mathbb{R}$, $\gamma_h : t \mapsto \gamma_h(t) = f(x_1^\star + th, \varphi(x_1^\star + th))$. Cette fonction présente un minimum local en 0 et comme elle est déribvable en 0, on a:

$$\dot{\gamma}_h(0) = 0 = (\partial_1 f(x^\star) + \partial_2 f(x^\star) \circ \varphi'(x_1^\star))(h)$$

$$(\partial_1 f(x^\star) + \partial_2 f(x^\star) \circ \varphi'(x_1^\star))(h) = 0$$

Comme $\varphi'(x_1^\star) = 0$, on en déduit que $\partial_1 f(x^\star) = 0$. Comme $h \in E_1$, on a donc

$$\partial_1 f(x^\star) = 0 = f'(x^\star)(h).$$

Nous avons bien donc $E_1 = ker(g'(x^\star)) \subset ker(f'(x^\star)) = \nabla f(x^\star)^\perp$.

• Conditions nécessaires du premier ordre (de Lagrange):

Théorème 13 *Soit $x^\star \in A$ une solution locale de (1). On suppose que:*

1- f est différentiable en $x^\star$

2- g est de classe C^1 au voisinage de $x^\star$

3- $\nabla g_1(x^\star), \ldots, \nabla g_m(x^\star)$ sont linéairement indépendants

alors il existe $(\lambda_1, \ldots, \lambda_m) \in I\!R^m$ tel que:

$$\nabla f(x^\star) = \sum_{j=1}^{m} \lambda_j \nabla g_j(x^\star) \tag{6}$$

Preuve: d'après la propriété 13, nous avons

$$\cap_{j=1}^{m} \nabla g_j(x^\star)^\perp \subset \nabla f(x^\star)^\perp$$

Il s'ensuit, d'après le lemme 4, l'existence de $(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$ vérifiant la relation (6).

Remarques 2 *1- Les réels $\lambda_1, \ldots, \lambda_m$ sont appelés des multiplicateurs de Lagrange associés aux contraintes de (1) au point de minimum $x^\star$.*

2- Les multiplicateurs de Lagrange sont uniques et ceci résulte du fait que $\nabla g_1(x^\star), \ldots, \nabla g_m(x^\star)$ sont linéairement indépendants.

3- Notons que l'hypothèse qui stipule que $\nabla g_1(x^\star), \ldots, \nabla g_m(x^\star)$ sont linéairement indépendants porte sur $x^\star$ qui est a priori inconnu! De même, l'indépendance de ces vecteurs implique que $m \leq n$, c'est-à-dire qu'il y a moins de contraintes que de variables.

• Cas convexe: condition nécessaire et suffisante:

Théorème 14 *Supposons que Ω un ouvert convexe de $\mathbb{R}^n$, que f une fonction convexe sur Ω et que g_j est une fonction affine pour $j = 1, \ldots, m$. Si $x^\star \in A$ est tel qu'il existe $(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$ tels que:*

$$\nabla f(x^\star) = \sum_{j=1}^{m} \lambda_j \nabla g_j(x^\star) \tag{7}$$

alors $f(x^\star) \leq f(x)$ pour tout $x \in A$.

Preuve: l'ensemble des contraines A peut s'écrire comme suit:
$A = \{x \in \Omega : g(x) = G.x = c\}$. Les fonction $g(x)$ tant affines, nous avons :

$$\forall x \in A : \; g_j(x) - g_j(x^\star) = \left(\nabla g_j(x^\star)\right)^T .(x - x^\star) = c_j - c_j = 0 \tag{8}$$

Par ailleurs, par convexité de f, on a: $f(x) - f(x^\star) \geq \left(\nabla f(x^\star)\right)^T .(x - x^\star)$

Avec (7) et (8), on obtient $f(x) - f(x^\star) \geq \sum_{j=1}^{m} \lambda_j \left(\nabla g_j(x^\star)\right)^T .(x - x^\star) = 0$

Lagrangien

Le lagrangien du problème (1) est la fonction définie sur $\Omega \times \mathbb{R}^m$ par:

$$\mathcal{L}(x, \lambda_1, \ldots, \lambda_m) = f(x) - \sum_{j=1}^{m} \lambda_j (g_j(x) - c_j) \tag{9}$$

Si la fonction f et les fonctions g_j sont différentiables en $x \in \Omega$, alors on a :

$$\nabla_x \mathcal{L}(x, \lambda_1, \ldots, \lambda_m) = \nabla f(x) - \sum_{j=1}^{m} \lambda_j \nabla g_j(x) \tag{10}$$

$$\nabla_{\lambda_j} \mathcal{L}(x, \lambda_1, \ldots, \lambda_m) = c_j - g_j(x), \quad j = 1, \ldots, m \tag{11}$$

Ainsi, comme on s'intéresse aux $x \in A$, et en posant $\lambda = (\lambda_1, \ldots, \lambda_m)$, les conditions (10) et (11) deviennent:

$$\nabla_x \mathcal{L}(x, \lambda_1, \ldots, \lambda_m) = 0 \tag{12}$$

$$\nabla_\lambda \mathcal{L}(x, \lambda_1, \ldots, \lambda_m) = 0 \tag{13}$$

Le théorème 13 peut se reformuler comme suit:

Théorème 15 *Soit $x^\star \in A$ une solution locale de (1). On suppose que:*

1- f est différentiable en $x^\star$

2- g est de classe C^1 au voisinage de $x^\star$

3- $\nabla g_1(x^\star), \ldots, \nabla g_m(x^\star)$ sont linéairement indépendants

alors il existe $(\lambda_1, \ldots, \lambda_m) \in I\!R^m$ tel que:

$$\mathcal{L}'(x, \lambda_1, \ldots, \lambda_m) = 0 \tag{14}$$

• Conditions nécessaires du second ordre :

Théorème 16 *Soit $x^\star \in A$ une solution locale de (1). On suppose que:*

1- f est deux fois différentiable en $x^\star$

2- g est de classe C^2 au voisinage de $x^\star$

3- $\nabla g_1(x^\star), \ldots, \nabla g_m(x^\star)$ sont linéairement indépendants

alors il existe $\lambda = (\lambda_1, \ldots, \lambda_m)^T \in I\!R^m$ tel que:

$$\begin{aligned} \nabla_x \mathcal{L}(x^\star, \lambda) = \nabla f(x^\star) - \sum_{j=1}^{m} \lambda_j \nabla g_j(x^\star) &= 0 \text{ et} \\ \partial^2_{xx} \mathcal{L}(x^\star, \lambda)(h,h) &\geq 0 \text{ pour tout } h \in ker(g'(x^\star)) \end{aligned}$$

• Conditions suffisantes du second ordre :

Théorème 17 *Soit $x^\star \in A$. On suppose que:*

1- f est deux fois différentiable en $x^\star$

2- g est de classe C^2 au voisinage de $x^\star$

3- $\nabla g_1(x^\star), \ldots, \nabla g_m(x^\star)$ sont linéairement indépendants

alors il existe $\lambda = (\lambda_1, \ldots, \lambda_m)^T \in I\!R^m$ tel que:

$$\nabla_x \mathcal{L}(x^\star, \lambda) = \nabla f(x^\star) - \sum_{j=1}^{m} \lambda_j \nabla g_j(x^\star) \quad = \quad 0 \text{ et}$$

$$\partial^2_{xx} \mathcal{L}(x^\star, \lambda)(h, h) \quad > \quad 0 \text{ pour tout } h \in ker(g'(x^\star)),\ h \neq 0$$

alors $x^\star$ est un minimum local strict de f sur A

Remarque 1 *Les conditions du second ordre (aussi bien nécessaires que suffisantes) ne peuvent être vérifiées qu'après la détermination des multiplicateurs de Lagrange $\lambda_1, \ldots, \lambda_m$.*

CHAPITRE 9

OPTIMISATION SOUS CONTRAINTES D'EGALITE ET D'INEGALITE

Formulation du problème

On s'intéresse à des problèmes d'optimisation sous contraintes d'égalité et d'inégalité dans $\mathbb{R}^n$.

• Ω un ouvert de $\mathbb{R}^n$, f, $g_1, \ldots, g_m$, $k_1, \ldots, k_p$ des fonctions définies sur Ω à valeurs dans $\mathbb{R}$, $(c_1, \ldots, c_n) \in \mathbb{R}^m$, $(d_1, \ldots, d_p) \in \mathbb{R}^p$.

Le problème considéré consiste à chercher

$$\inf_{x \in A} f(x) \tag{15}$$

avec

$$A = \{x \in \Omega : \ g_j(x) = c_j, \ j = 1, \ldots, m, \ k_i(x) \leq d_i, i = 1, \ldots, p\}$$

La fonction f est appelée fonction objectif ou fonction coût. Les fonctions g_j et

les réels c_j définissent les contraintes d'égalité. Les fonctions k_i et les réels d_i définissent les contraintes d'inégalité. Les éléments de A s'appellent les éléments admissibles.

Quelques définitions et notations

On introduit la notation suivante $g : \Omega \to \mathbb{R}^m, x \mapsto g(x) = (g_1(x), \ldots, g_m(x))^T$. Ainsi, les contraines d'éégalité peuvent s'écrire comme suit: $g(x) = c$ où $c = (c_1, \ldots, c_m)^T \in \mathbb{R}^m$.

Soit maintenant $x \in A$. Si $k_i(x) = d_i$ alors on dit que la $i - ème$ contrainte d'inégalité est saturée en x (ou encore est active en x). Désignons maintenant par $I(x)$ l'ensemble des indices correspondant aux contraintes saturées en x, c'est-à-dire:

$$I(x) = \{i \in \{1, \ldots, p\} \text{ tel que } k_i(x) = d_i\}$$

Préliminaires

Dans tout ce qui suit, on considère $x^\star \in A$ et on suppose que les conditions suivantes sont satisfaites:

1- g est de classe C^1 au voisinage de $x^\star$,

2- k_i est différentiable en $x^\star$ pour tout $i \in I(x^\star)$,

3- $\nabla g_1(x^\star), \ldots, \nabla g_m(x^\star)$ sont linéairement indépendants,

4- Les contraintes d'inégalité sont qualifiées en $x^\star$, c'est-à-dire

$$\exists h_0 \in ker(g'(x^\star)) \text{ tel que } \left(\nabla k_i(x^\star)\right)^T .h_0 < 0 \ \forall i \in I(x^\star) \tag{16}$$

Notons que l'hypothèse de qualification (21) peut s'exprimer de manière équivalente comme suit

$$\begin{aligned} \exists h_0 \text{ tel que } \left(\nabla g_j(x^\star)\right)^T .h_0 &= 0 \ , j = 1, \ldots, m, \text{ et} \\ \left(\nabla k_i(x^\star)\right)^T .h_0 &< 0 \ \forall i \in I(x^\star) \end{aligned} \tag{17}$$

Conditions d'optimalité de K.K.T.

Nous allons énoncer le théorème de Karush, Kuhn et Tucker (KKT).

Théorème 18 *Soit $x^\star \in A$ une solution locale de (15) telle que*

1- f est différentiable en $x^\star$,

2- g est de classe C^1 au voisinage de $x^\star$,

3- k_i est différentiable en $x^\star$ pour tout $i \in I(x^\star)$,

4- $\nabla g_1(x^\star), \ldots, \nabla g_m(x^\star)$ sont linéairement indépendants,

5- Les contraintes d'inégalité sont qualifiées en $x^\star$, c'est-à-dire

$$\exists h_0 \in ker(g'(x^\star)) \text{ tel que } \left(\nabla k_i(x^\star)\right)^T .h_0 < 0 \ \forall i \in I(x^\star) \tag{18}$$

alors il existe $(\lambda_1, \ldots, \lambda_m) \in I\!R^m$ et $\mu_i \geq 0$ pour tout $i \in I(x^\star)$ tel que

$$\nabla f(x^\star) = \sum_{j=1}^{m} \lambda_j \nabla g_j(x^\star) - \sum_{i \in I(x^\star)} \mu_i \nabla k_i(x^\star) \tag{19}$$

Cas convexe

Lorsque le problème (15) est convexe, la condition (19) devient suffisante et assure que le minimum $x^\star$ est global.

Théorème 19 *Supposons que Ω est un ouvert convexe de $\mathbb{R}^n$, que f et $k_1, \ldots, k_p$ sont des fonctions convexes sur Ω, g_j est une fonction affine pour $j = 1, \ldots, m$. Si $x^\star \in A$ est tel qu'il existe $(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$ et $\mu_i \geq 0$ pour tout $i \in I(x^\star)$ tel que:*

$$\nabla f(x^\star) = \sum_{j=1}^{m} \lambda_j \nabla g_j(x^\star) - \sum_{i \in I(x^\star)} \mu_i \nabla k_i(x^\star)$$

alors $f(x^\star) \leq f(x)$ pour tout $x \in A$.

Lagrangien

Les conditions d'optimalité KKT peuvent s'exprimer au moyen du Lagrangien associé au problème (15). L'idée est d'introduire des multiplicateurs , $\mu_i, i = 1, \ldots, p$, pour toutes les contraintes d'inégalité:

- $\mu_i = 0$ si $i \notin I(x^\star)$
- $\mu_i > 0$ si $i \in I(x^\star)$, c'est-à-dire si $k_i(x^\star) - d_i = 0$.

De ce fait, si une contrainte de type inégalité n'est pas saturée, le multiplicateur associé est nul. On aura ainsi

$$\mu_i(k_i(x^\star) - d_i) = 0 \text{ pour tout } i =, 1, \ldots, p$$

Lagrangien

Le lagrangien du problème (15) est la fonction définie pour tout $(x, \lambda, \mu) \in \Omega \times \mathbb{R}^m \times (\mathbb{R}_+)^p$ par:

$$\mathcal{L}(x, \lambda, \mu) = f(x) - \sum_{j=1}^{m} \lambda_j (g_j(x) - c_j) + \sum_{i=1}^{m} \mu_i (k_i(x) - d_i) \tag{20}$$

Si $f, g_1, \ldots g_m, k_1, \ldots, k_p$ sont différentiables en $x \in \Omega$, on a:

$$\begin{aligned}
\nabla_x \mathcal{L}(x, \lambda, \mu) &= \nabla f(x) - \sum_{j=1}^{m} \lambda_j \nabla g_j(x) + \sum_{i=1}^{m} \mu_i \nabla k_i(x) \\
\nabla_{\lambda_j} \mathcal{L}(x, \lambda, \mu) &= c_j - g_j(x) \\
\nabla_{\mu_i} \mathcal{L}(x, \lambda, \mu) &= k_i(x) - d_i
\end{aligned}$$

Le théorème 18 peut se reformuler ainsi

Théorème 20 *Soit $x^\star \in A$ une solution locale de (15) telle que*

1- f est différentiable en $x^\star$,

2- g est de classe C^1 au voisinage de $x^\star$,

3- k_i est différentiable en $x^\star$ pour tout $i \in I(x^\star)$,

4- $\nabla g_1(x^\star), \ldots, \nabla g_m(x^\star)$ sont linéairement indépendants,

5- Les contraintes d'inégalité sont qualifiées en $x^\star$, c'est-à-dire

$$\exists h_0 \in ker(g'(x^\star)) \text{ tel que } \left(\nabla k_i(x^\star)\right)^T .h_0 < 0 \ \forall i \in I(x^\star) \tag{21}$$

alors il existe $(\lambda_1, \ldots, \lambda_m) \in I\!R^m$ et $(\mu_1, \ldots, \mu_p) \in (I\!R_+)^p$ tels que

$$\begin{aligned} \nabla_x \mathcal{L}(x^\star, \lambda, \mu) &= 0 \\ \nabla_\lambda \mathcal{L}(x^\star, \lambda, \mu) &= 0 \\ \mu_i(k_i(x^\star) - d_i) &= 0 \ i = 1, \ldots, p \end{aligned}$$

Bibliographie

- Culioli, J.C. (1994). Introduction à l'optimisation, Ellipses.

- W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery (1992). Numerical Recipes in C, the Art of Scientific Computing, Cambridge University Press.

- Support de cours de l'IUP Génie Mathématique et Informatique de l'Université Paris Dauphine.